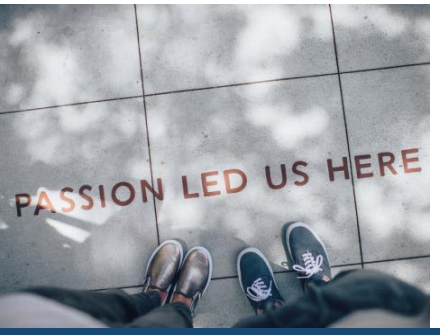


Using Big Data to Solve Economic and Social Problems

Professor Raj Chetty
Head Section Leader: Gregory Bruich, Ph.D.
Harvard University

Spring 2019

HARVARD UNIVERSITY

OPPORTUNITY INSIGHTS

Improving Health Outcomes

- Research in economics typically focuses on earnings or wealth as key outcomes of interest

- But most people view health and life expectancy as among the most important aspects of well-being

- What interventions are most effective in improving health (holding fixed current frontier of medical technology)?

    - Research on these issues spans multiple fields, from epidemiology and public health to economics

Improving Health Outcomes: Overview

- This part of the class illustrates how big data is helping us learn how to improve health, in three segments:

1. Descriptive analysis of health outcomes in U.S. population
   [method: survival analysis]

   Chetty, Stepner, Abraham, Lin, Scuderi, Bergeron, Cutler. “The Association Between Income and Life Expectancy in the United States” *JAMA* 2016.

Improving Health Outcomes: Overview

- This part of the class illustrates how big data is helping us learn how to improve health, in three segments:

1. Descriptive analysis of health outcomes in U.S. population [method: survival analysis]

2. Economics applications: impacts of food stamps (Jesse Shapiro) and health insurance [method: regression discontinuities]

   Wherry, Miller, Kaestner, Meyer. “Childhood Medicaid Coverage and Later Life Health Care Utilization” *REStat* 2017.

Improving Health Outcomes: Overview

- This part of the class illustrates how big data is helping us learn how to improve health, in three segments:

1. Descriptive analysis of health outcomes in U.S. population [method: survival analysis]

2. Economics applications: impacts of food stamps (Jesse Shapiro) and health insurance [method: regression discontinuities]

3. Epidemiology application: using big data to forecast pandemics [method: predictive modeling]

Ginsberg, Mohebbi, Patel, Brammer, Smolinski, Brilliant. “Detecting Influenza Epidemics Using Search Engine Query Data.” *Nature* 2009.

Lazer, Kennedy, King, Vespignani. “The Parable of Google Flu: Traps in Big Data Analysis.” *Science* 2014.

Income and Life Expectancy

- Most common measure of health: mortality rates
  - Crude but well measured in population data
- Begin with basic descriptive facts about life expectancy in America
- Chetty et al. (2016) examine relationship between life expectancy and income
  - Use data on entire U.S. population from 1999-2013 (1.4 billion observations)

Estimating Life Expectancy: Data

- Mortality measured using Social Security death records
- Income measured at household level using tax returns
- Focus on percentile ranks in income distribution
  - Rank individuals in national income distribution within birth cohort, gender, and tax year

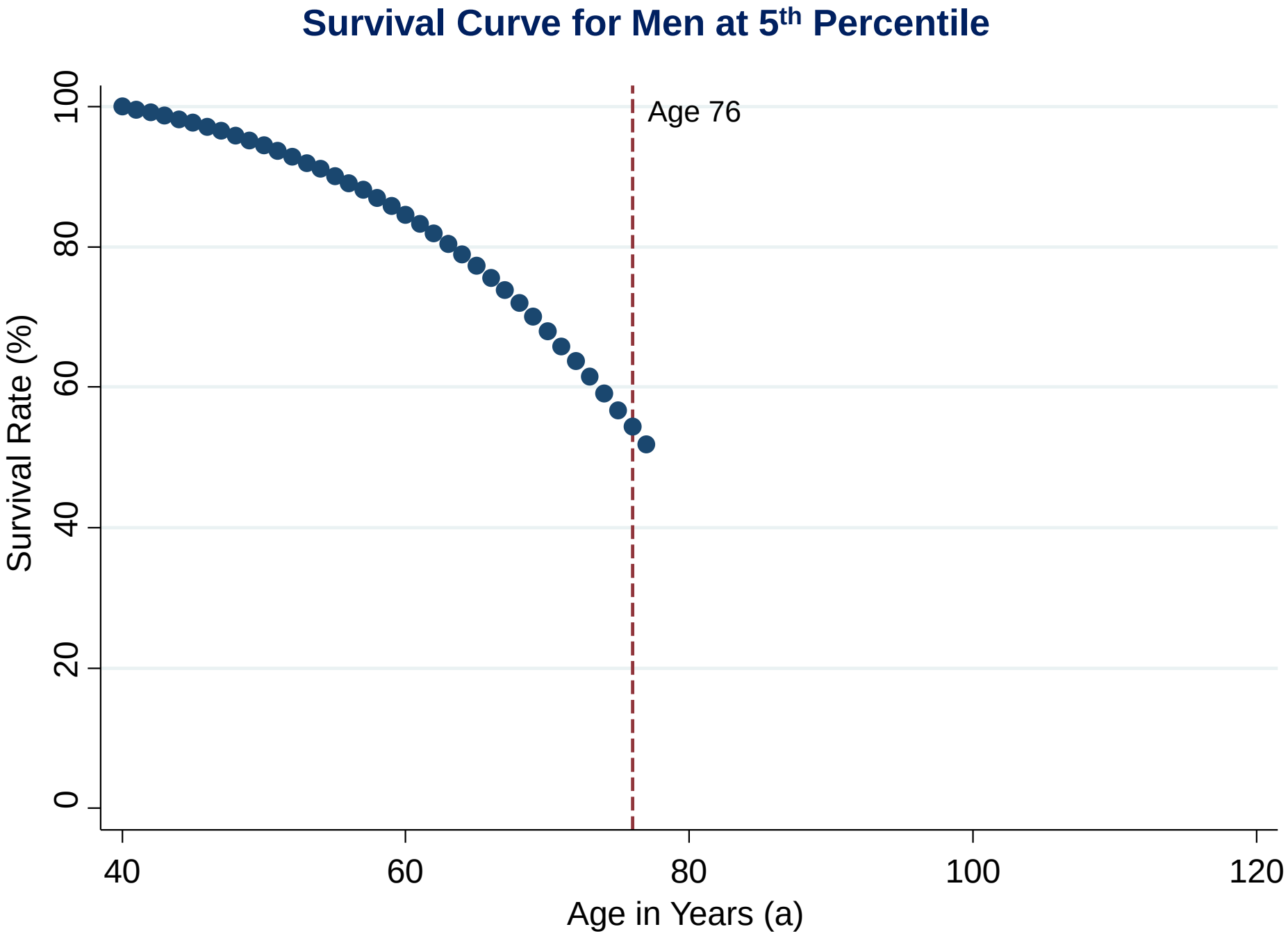
Methodology to Estimate Life Expectancy

- Goal: estimate expected age of death conditional on an individual's income at age 40, controlling for differences in race and ethnicity

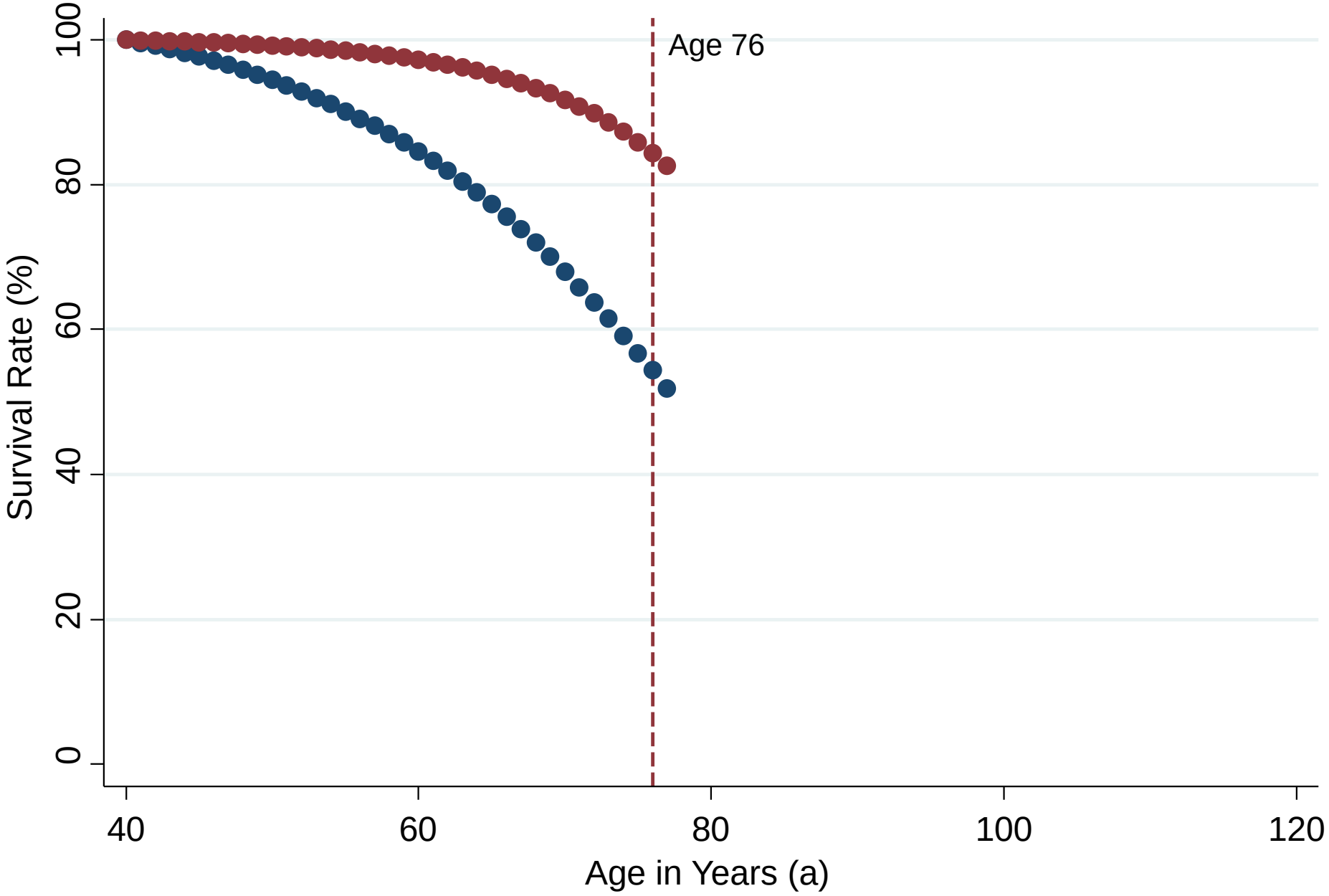
    - *Period* life expectancy: life expectancy for a hypothetical individual who experiences mortality rates at each age observed in a given year

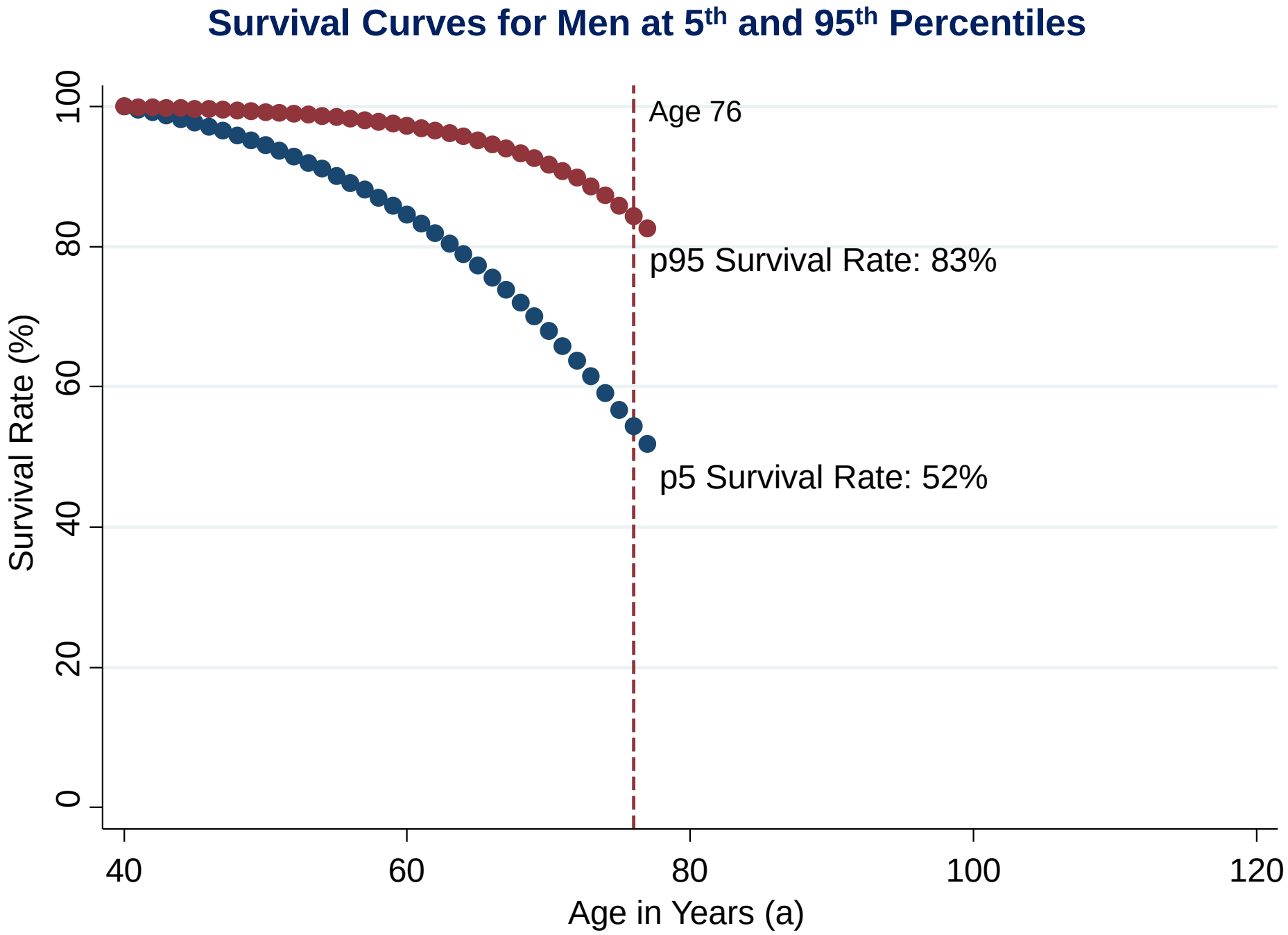
- Three steps:

    1. Calculate mortality rates by income rank and age for observed ages
    2. Estimate a survival model to extrapolate to older ages
    3. Adjust for racial differences in mortality rates


Survival Curve for Men at 5th Percentile
Age 76
Survival Rate (%)
0
20
40
60
80
100
40
60
80
100
120
Age in Years (a)

Survival Curves for Men at 5th and 95th Percentiles




Survival Curves for Men at 5th and 95th Percentiles
Survival Rate (%)
0
20
40
60
80
100
Age 76
p95 Survival Rate: 83%
p5 Survival Rate: 52%
40
60
80
100
120
Age in Years (a)

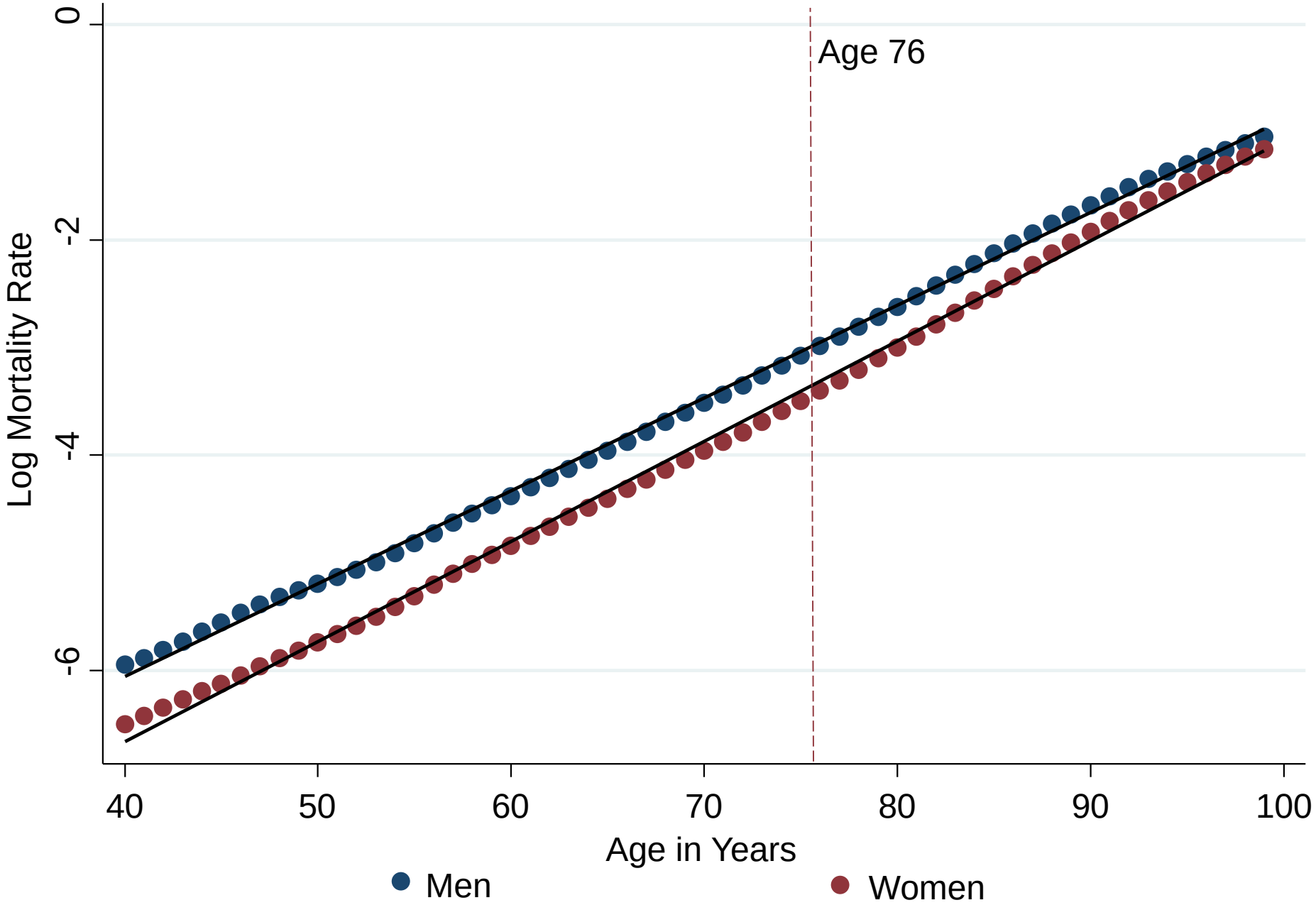
Step 2: Predicting Mortality Rates at Older Ages

- To calculate life expectancy, need estimates of mortality rates beyond age 76

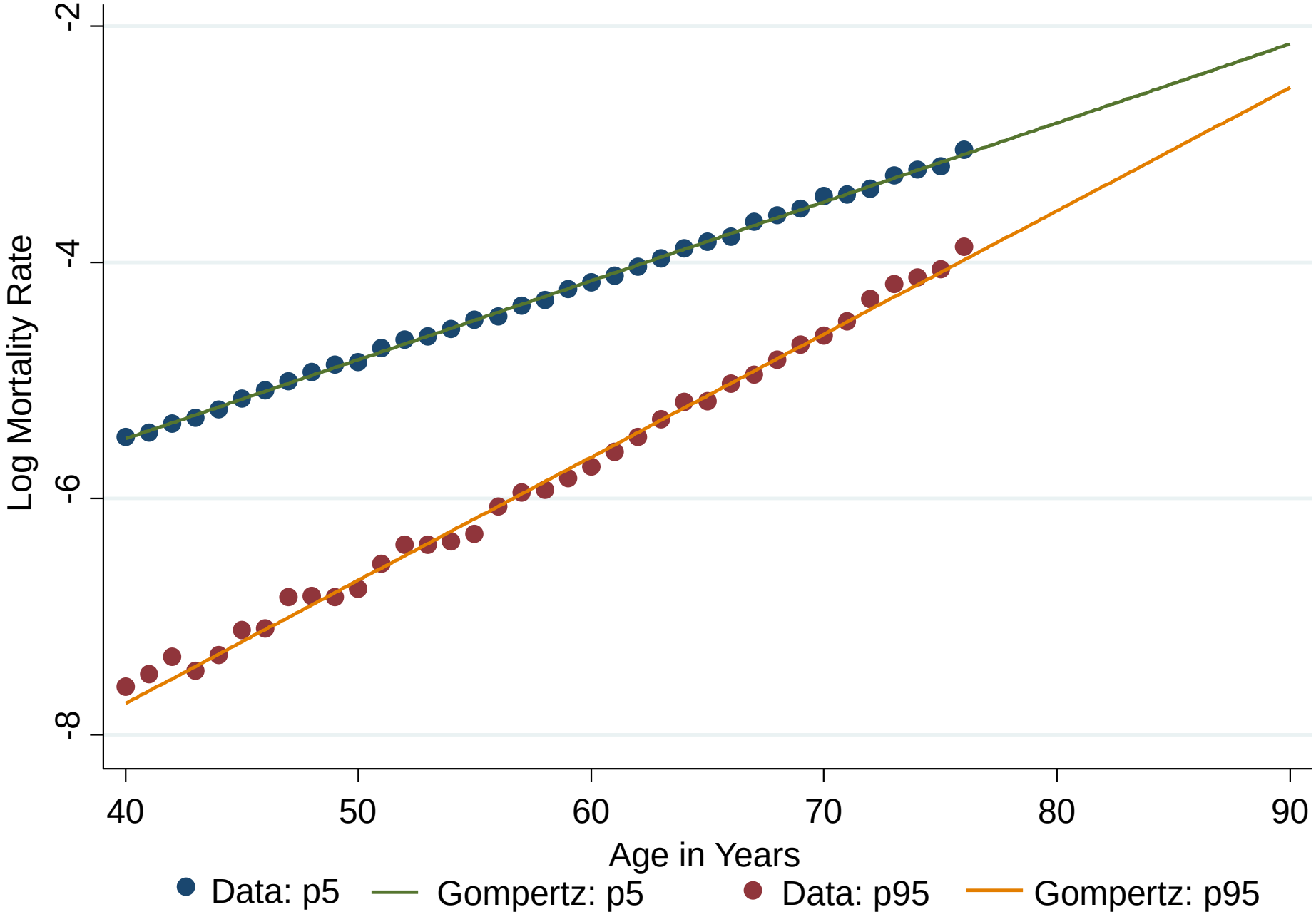
- Gompertz (1825) documented a robust empirical pattern: mortality rates grow exponentially with age

$$m(a) = k\, e^{\beta a}$$
$$\Rightarrow \log m(a) = \kappa + \beta a$$

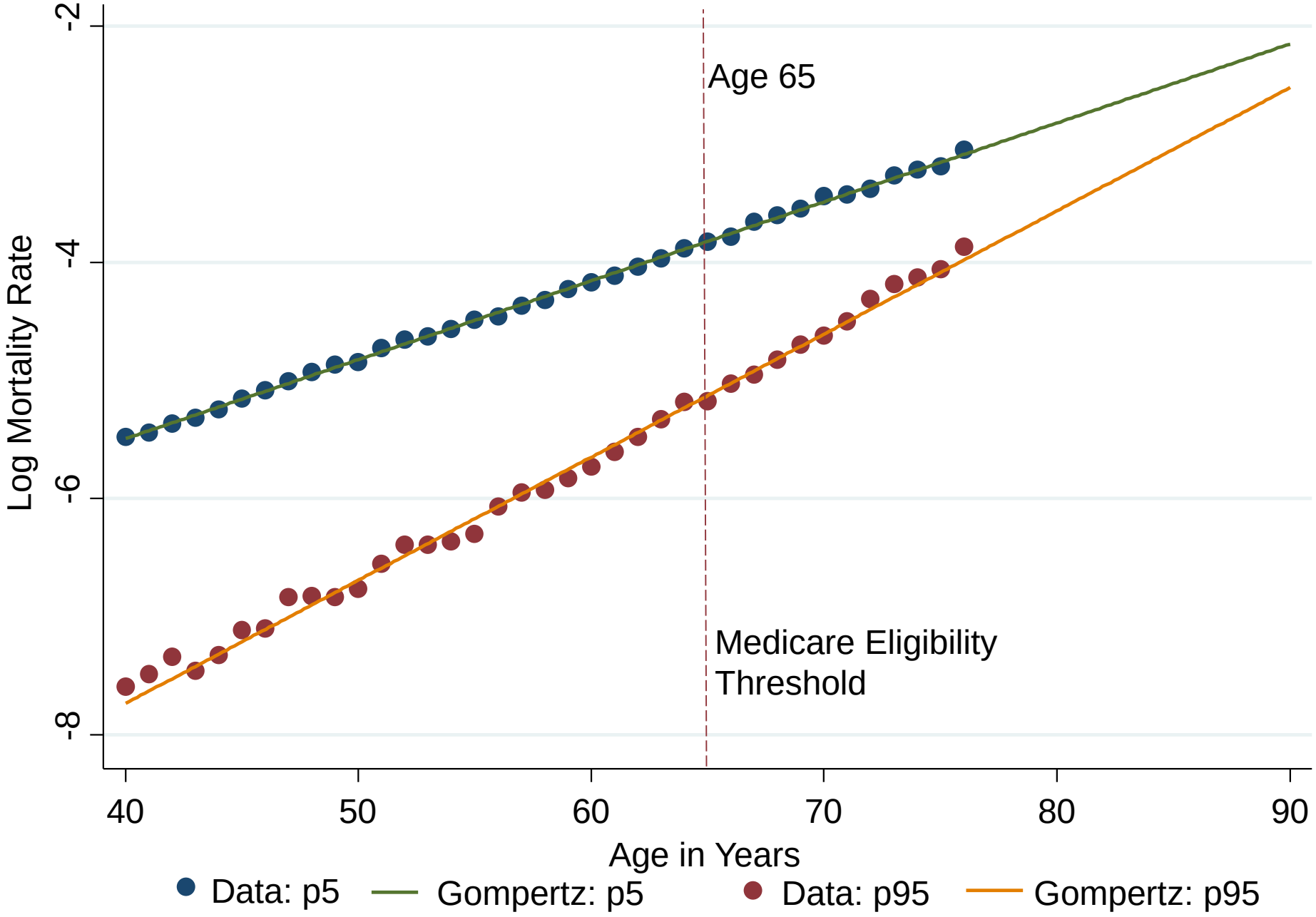
Mortality Rates by Gender in the United States in 2001: CDC Data

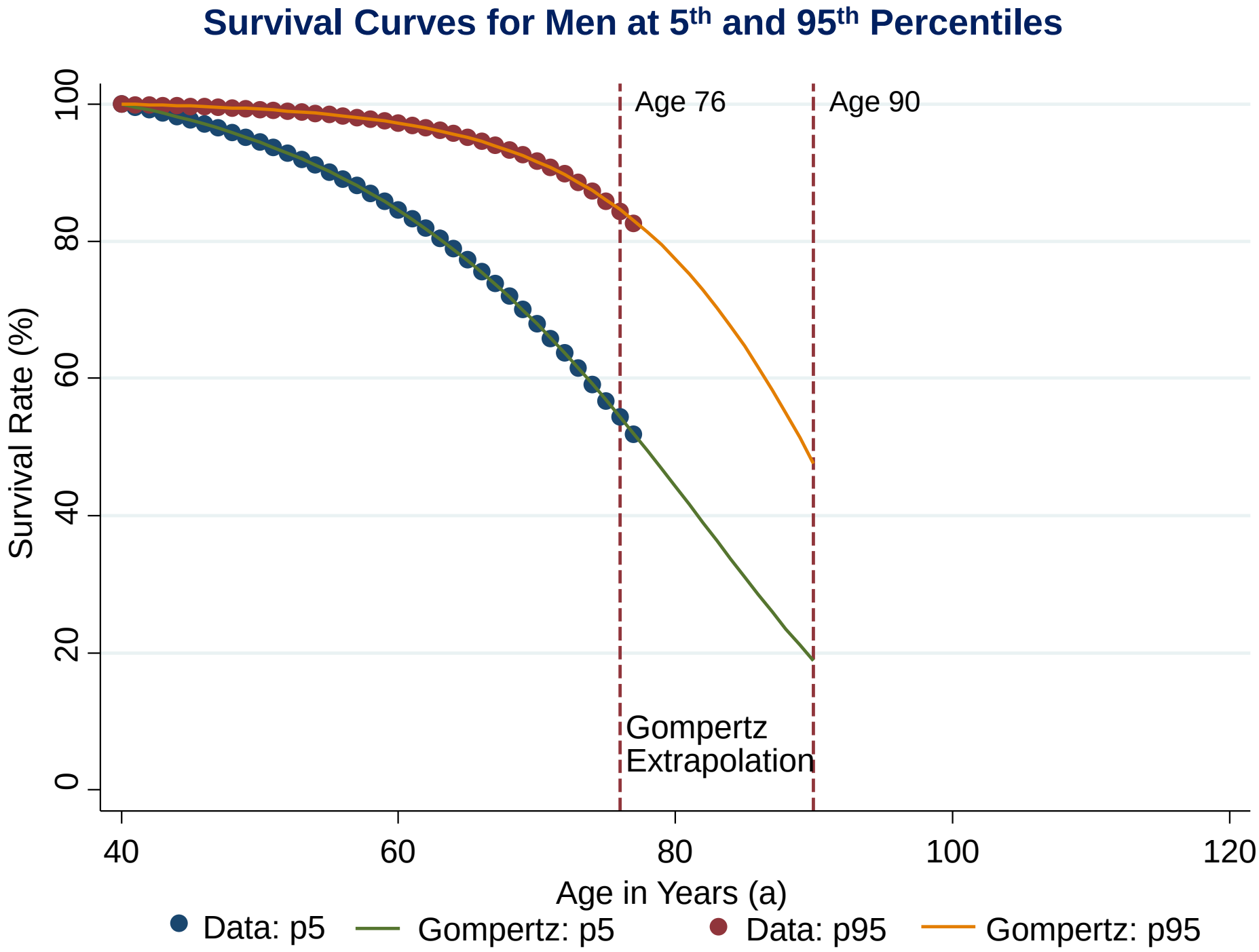


Log Mortality Rates for Men at 5th and 95th Percentiles

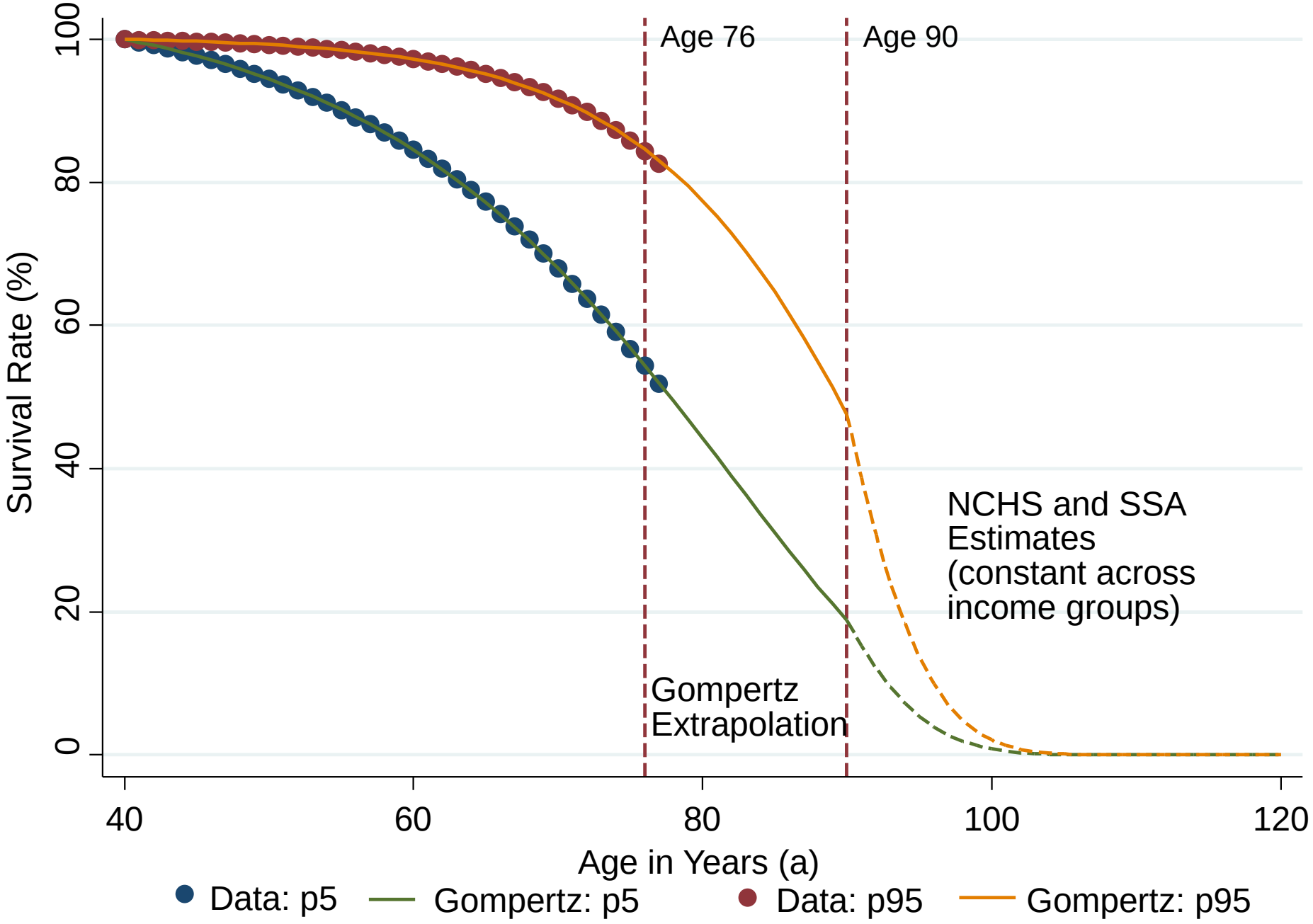


Log Mortality Rates for Men at 5th and 95th Percentiles

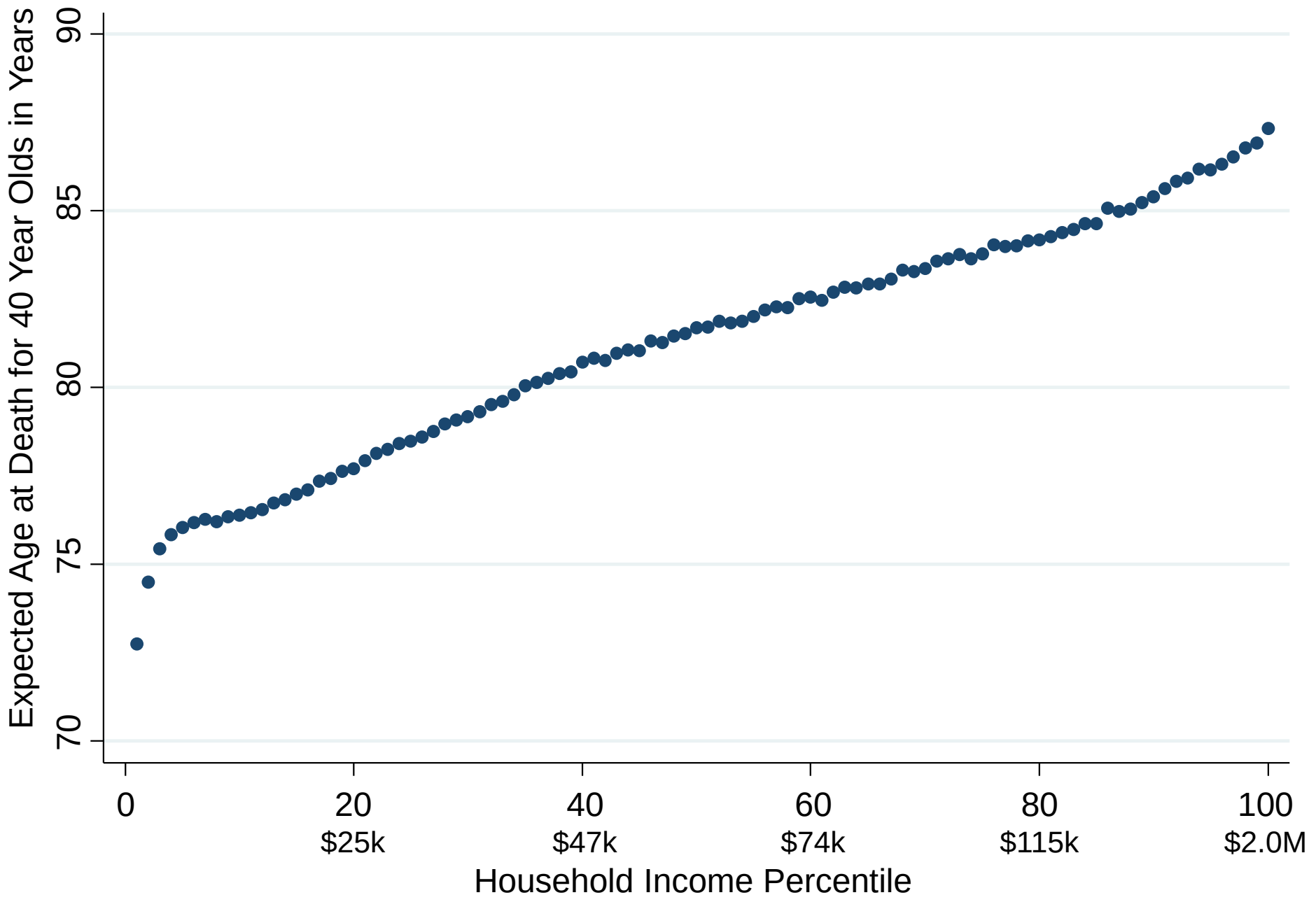


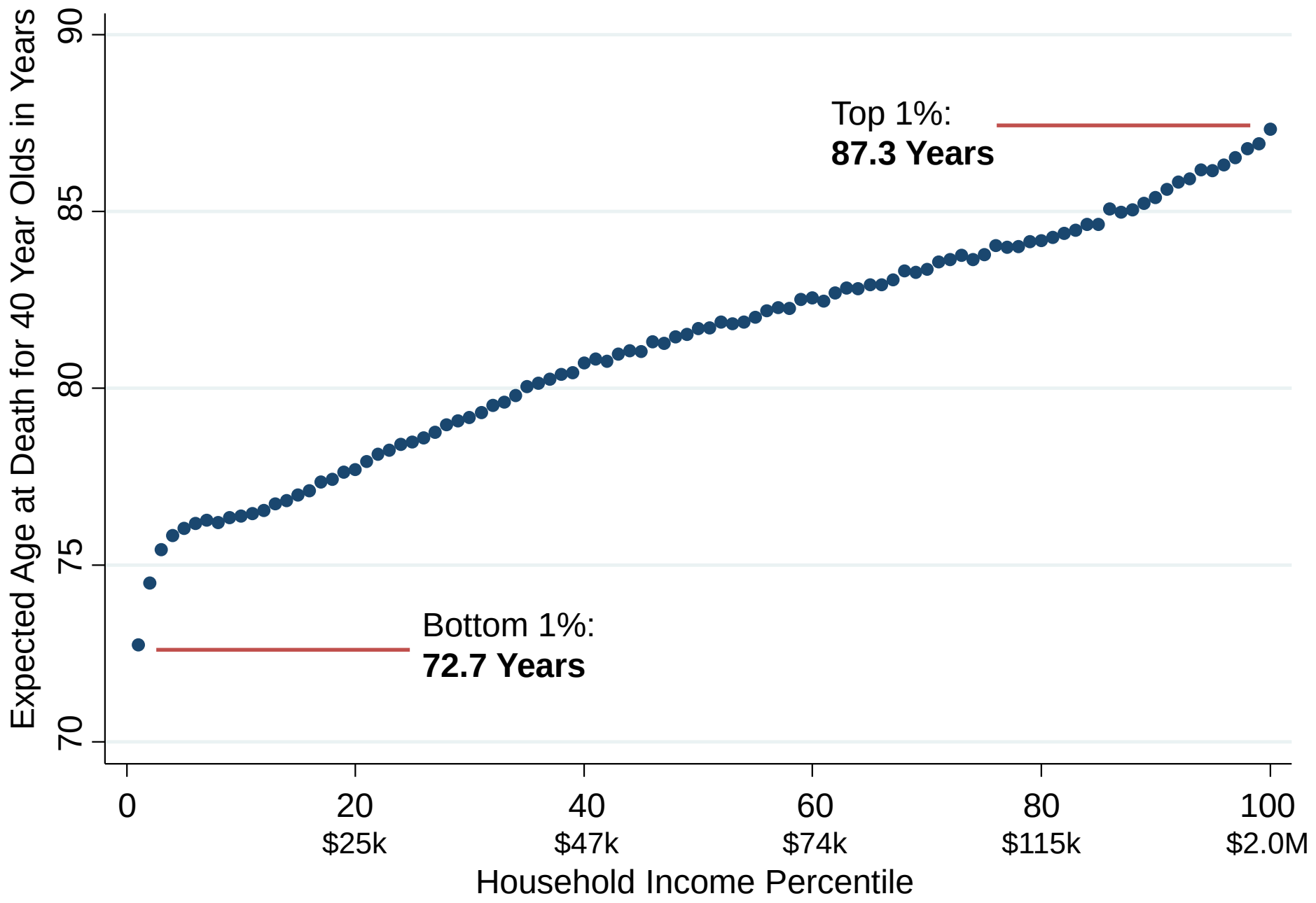

Survival Curves for Men at 5th and 95th Percentiles
Survival Rate (%)
0
20
40
60
80
100
Age 76
Age 90
Gompertz
Extrapolation
40
60
80
100
120
Age in Years (a)
Data: p5
Gompertz: p5
Data: p95
Gompertz: p95

Survival Curves for Men at 5th and 95th Percentiles

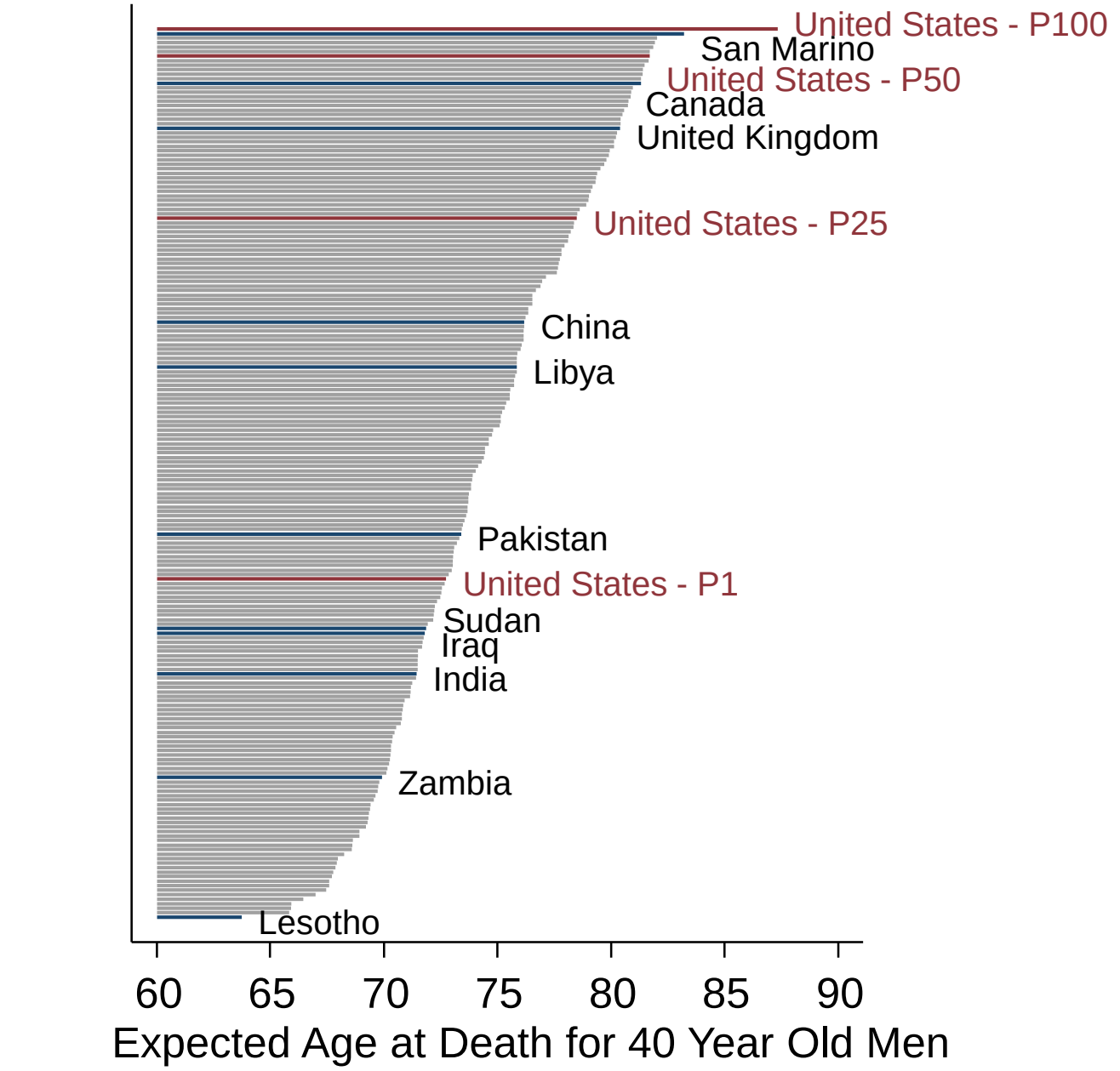


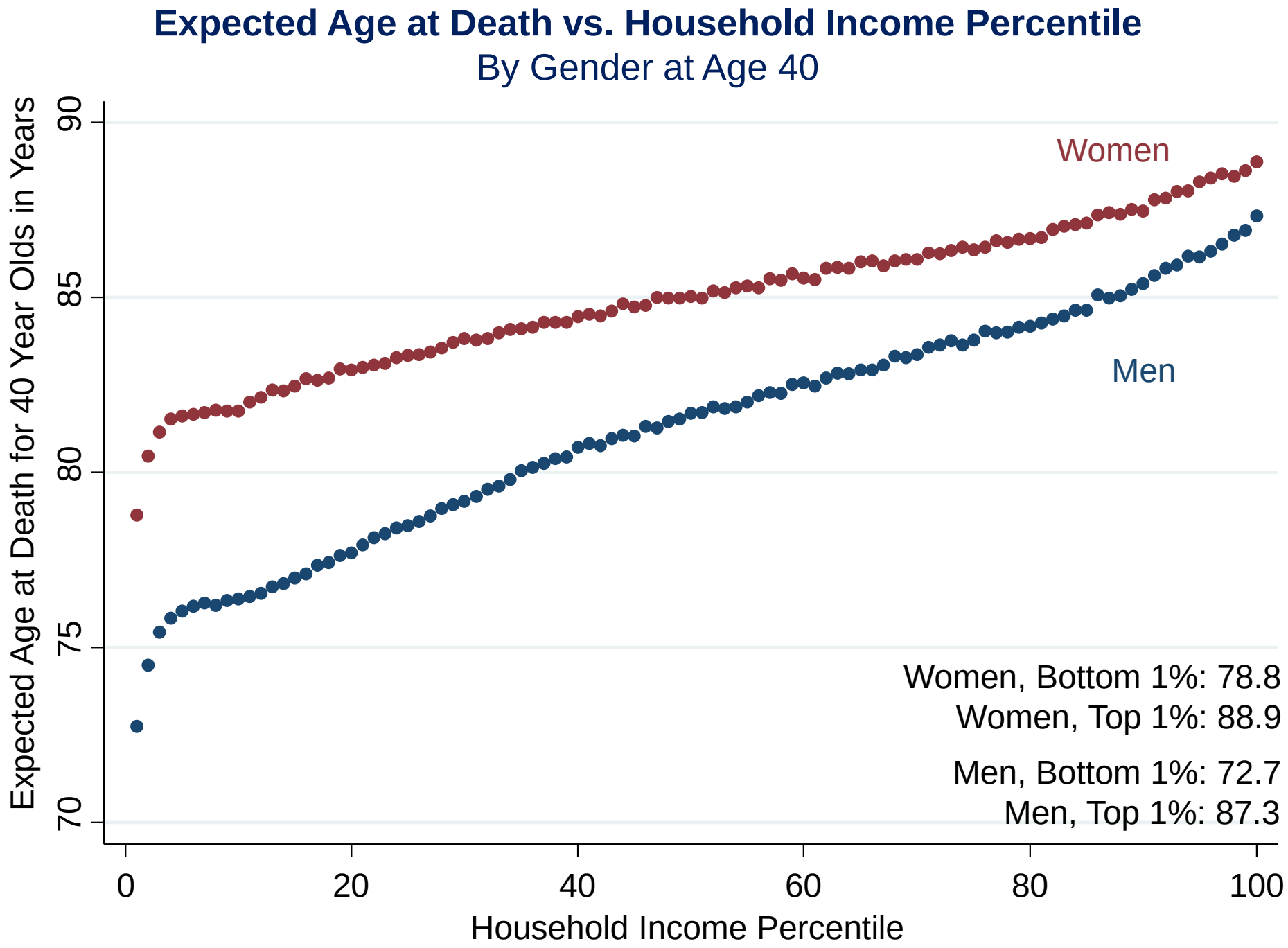
National Statistics on Income and Life Expectancy


Expected Age at Death vs. Household Income Percentile
For Men at Age 40
Expected Age at Death for 40 Year Olds in Years
70
75
80
85
90
0
20
40
60
80
100
$25k
$47k
$74k
$115k
$2.0M
Household Income Percentile


Expected Age at Death vs. Household Income Percentile
For Men at Age 40
Expected Age at Death for 40 Year Olds in Years
70
75
80
85
90
Top 1%:
87.3 Years
Bottom 1%:
72.7 Years
0
20
$25k
40
$47k
60
$74k
80
$115k
100
$2.0M
Household Income Percentile

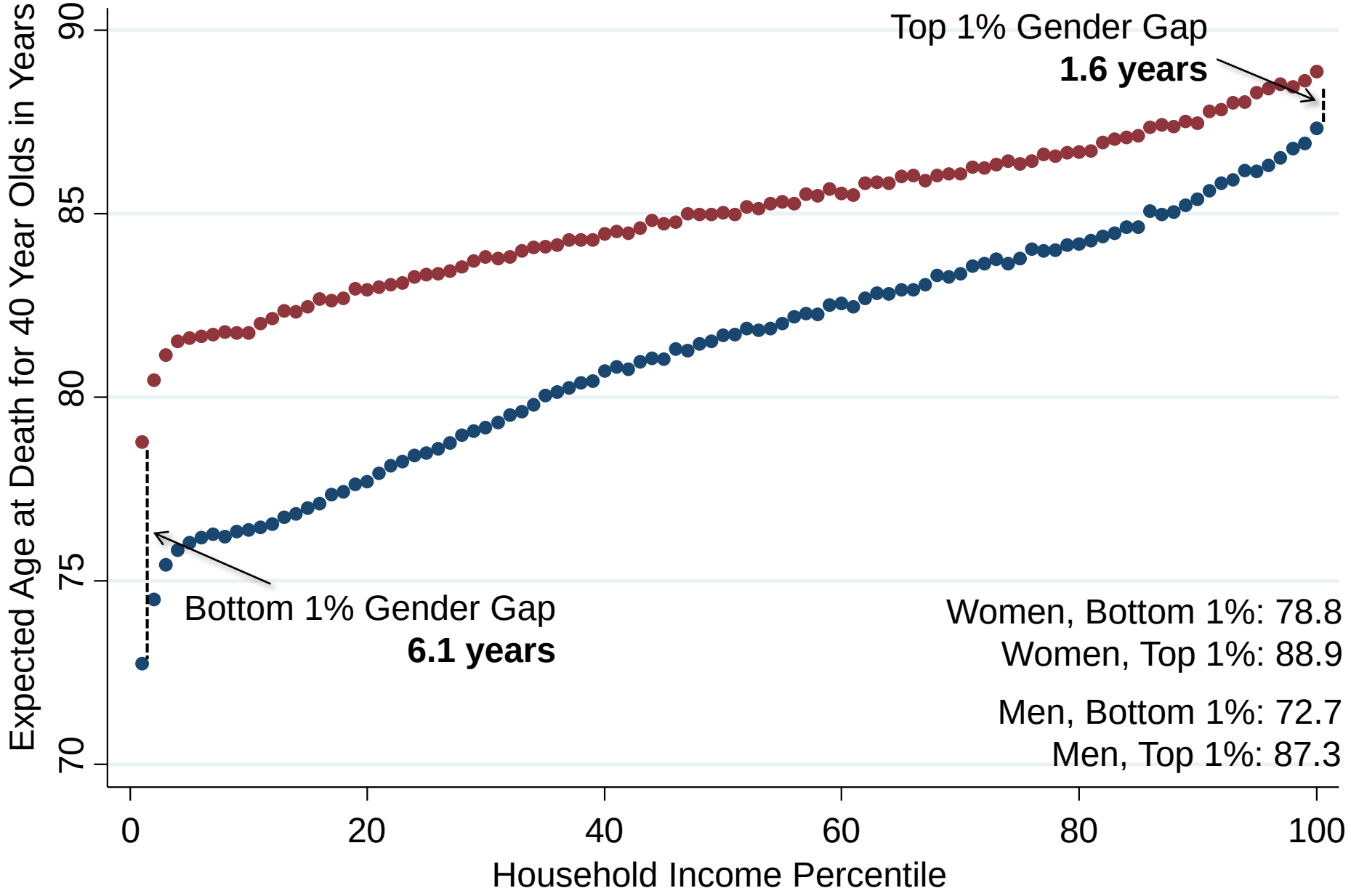
U.S. Life Expectancies by Percentile in Comparison to Mean Life Expectancies Across Countries




Expected Age at Death vs. Household Income Percentile
By Gender at Age 40
Expected Age at Death for 40 Year Olds in Years
70
75
80
85
90
Women
Men
Women, Bottom 1%: 78.8
Women, Top 1%: 88.9
Men, Bottom 1%: 72.7
Men, Top 1%: 87.3
0
20
40
60
80
100
Household Income Percentile

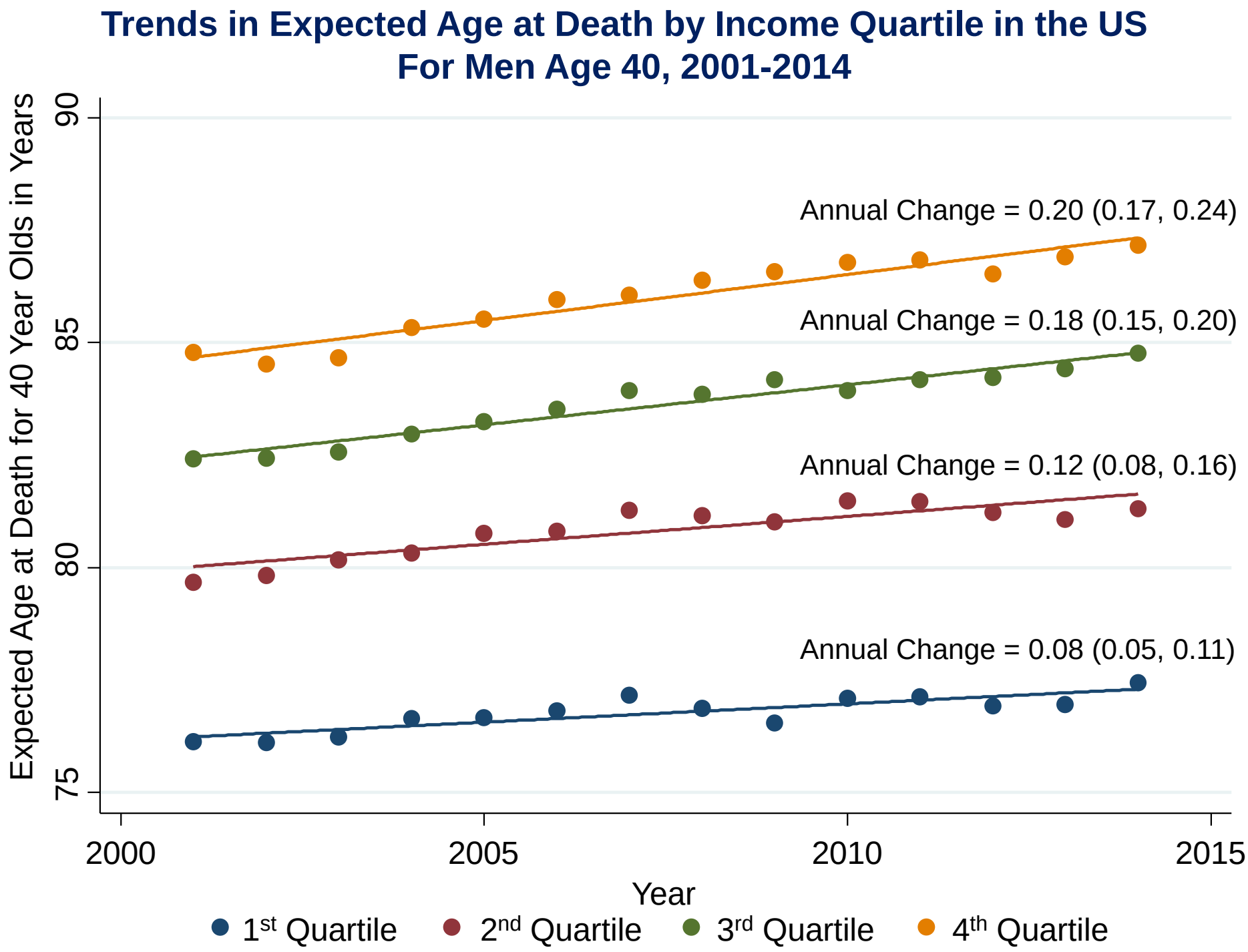
Expected Age at Death vs. Household Income Percentile

By Gender at Age 40

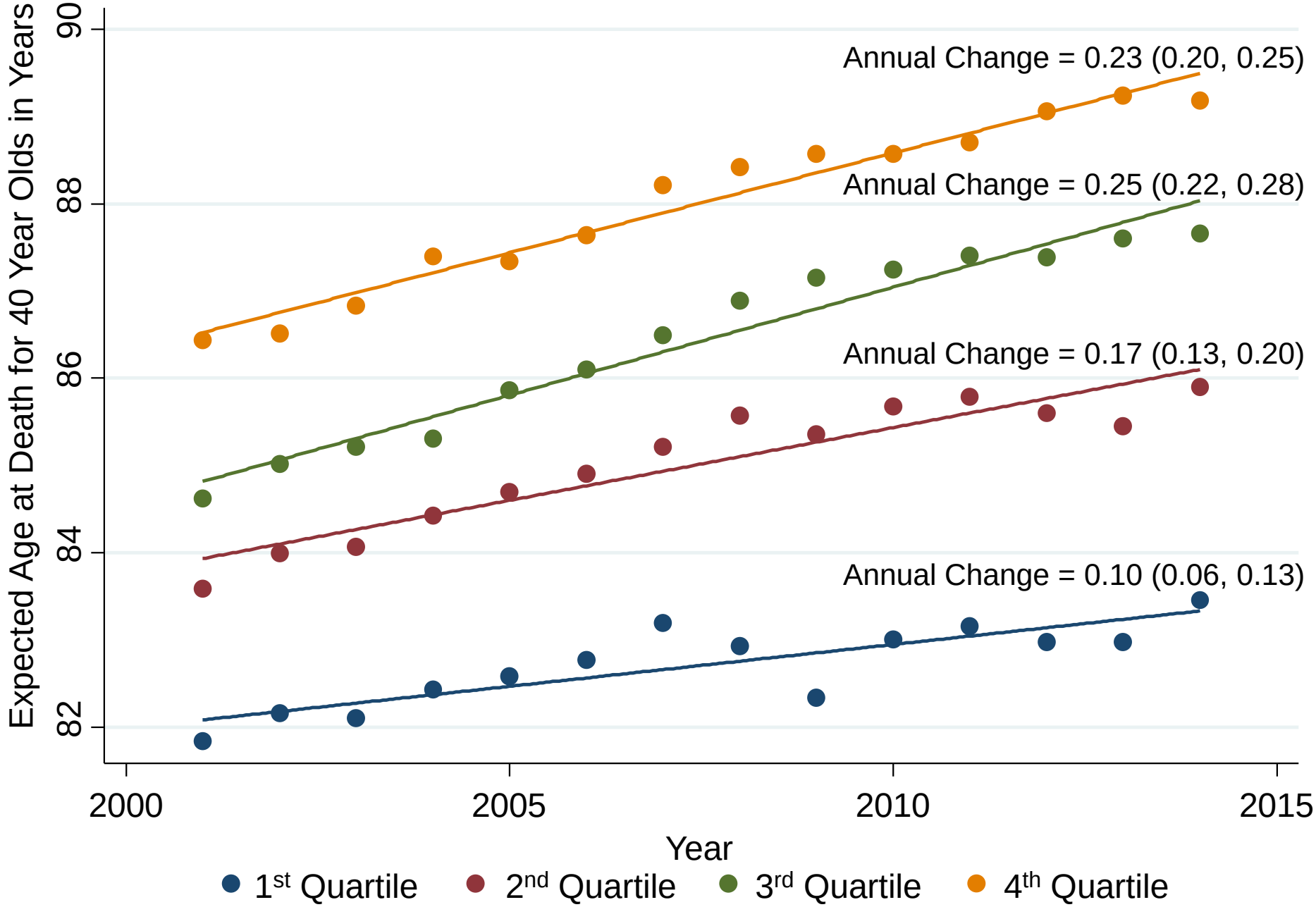


Time Trends

- How are gaps in life expectancy changing over time?

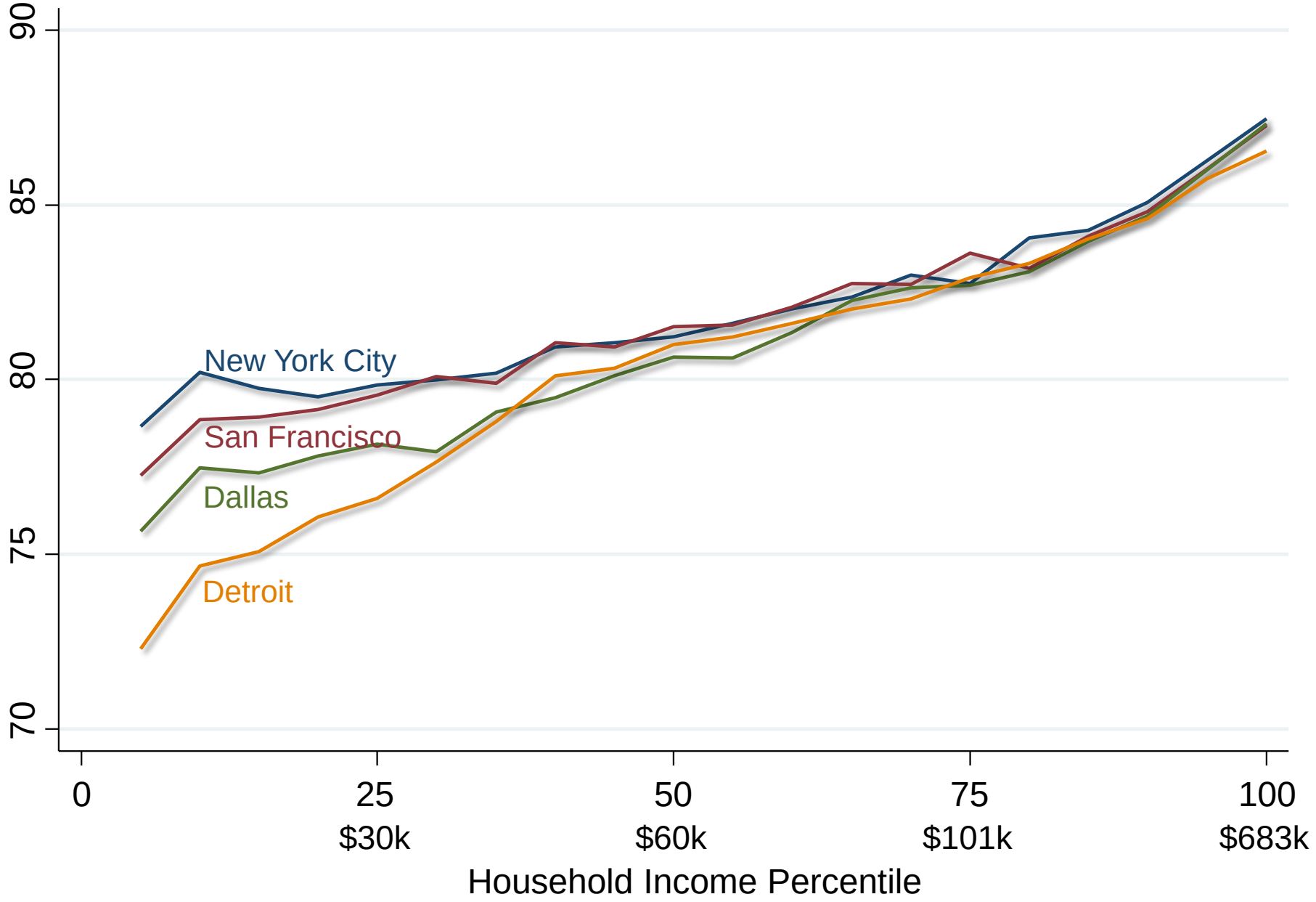

Trends in Expected Age at Death by Income Quartile in the US
For Men Age 40, 2001-2014
Expected Age at Death for 40 Year Olds in Years
75
80
85
90
2000
2005
2010
2015
Year
Annual Change = 0.20 (0.17, 0.24)
Annual Change = 0.18 (0.15, 0.20)
Annual Change = 0.12 (0.08, 0.16)
Annual Change = 0.08 (0.05, 0.11)
1st Quartile
2nd Quartile
3rd Quartile
4th Quartile

Trends in Expected Age at Death by Income Quartile in the US For Women Age 40, 2001-2014

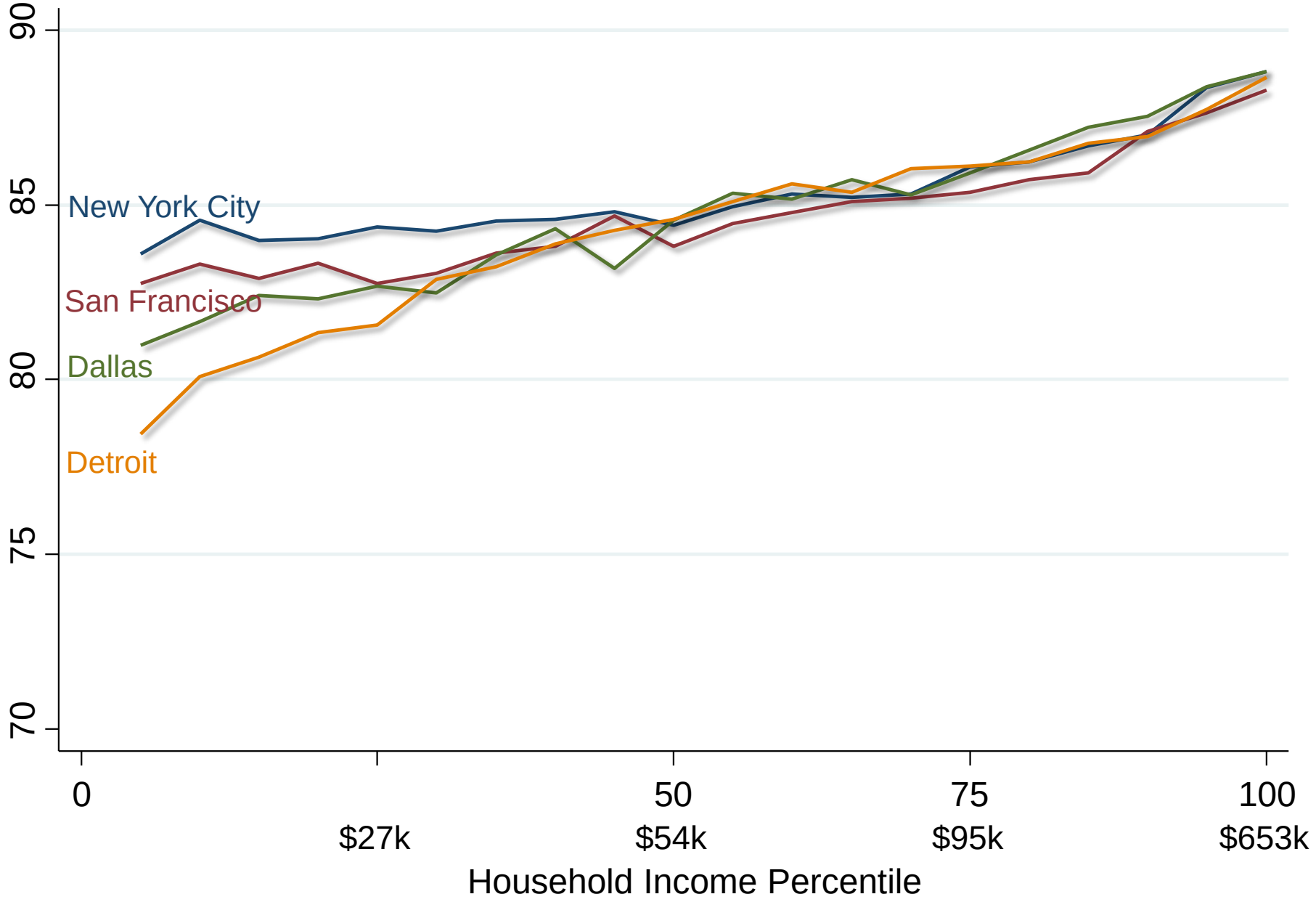


Local Area Variation in Life Expectancy by Income

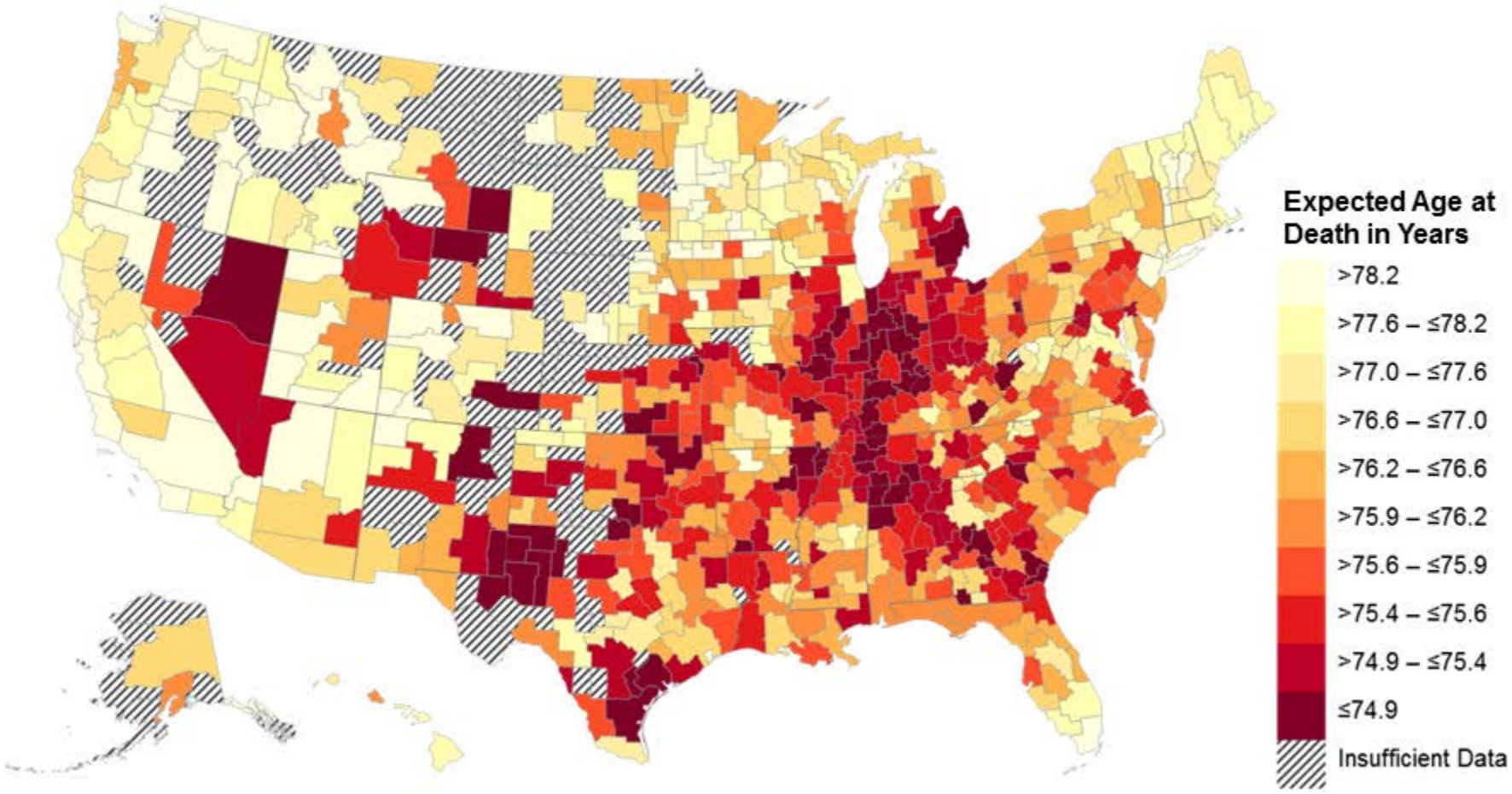
Expected Age at Death vs. Household Income for Men in Selected Cities



Expected Age at Death vs. Household Income for Women in Selected Cities

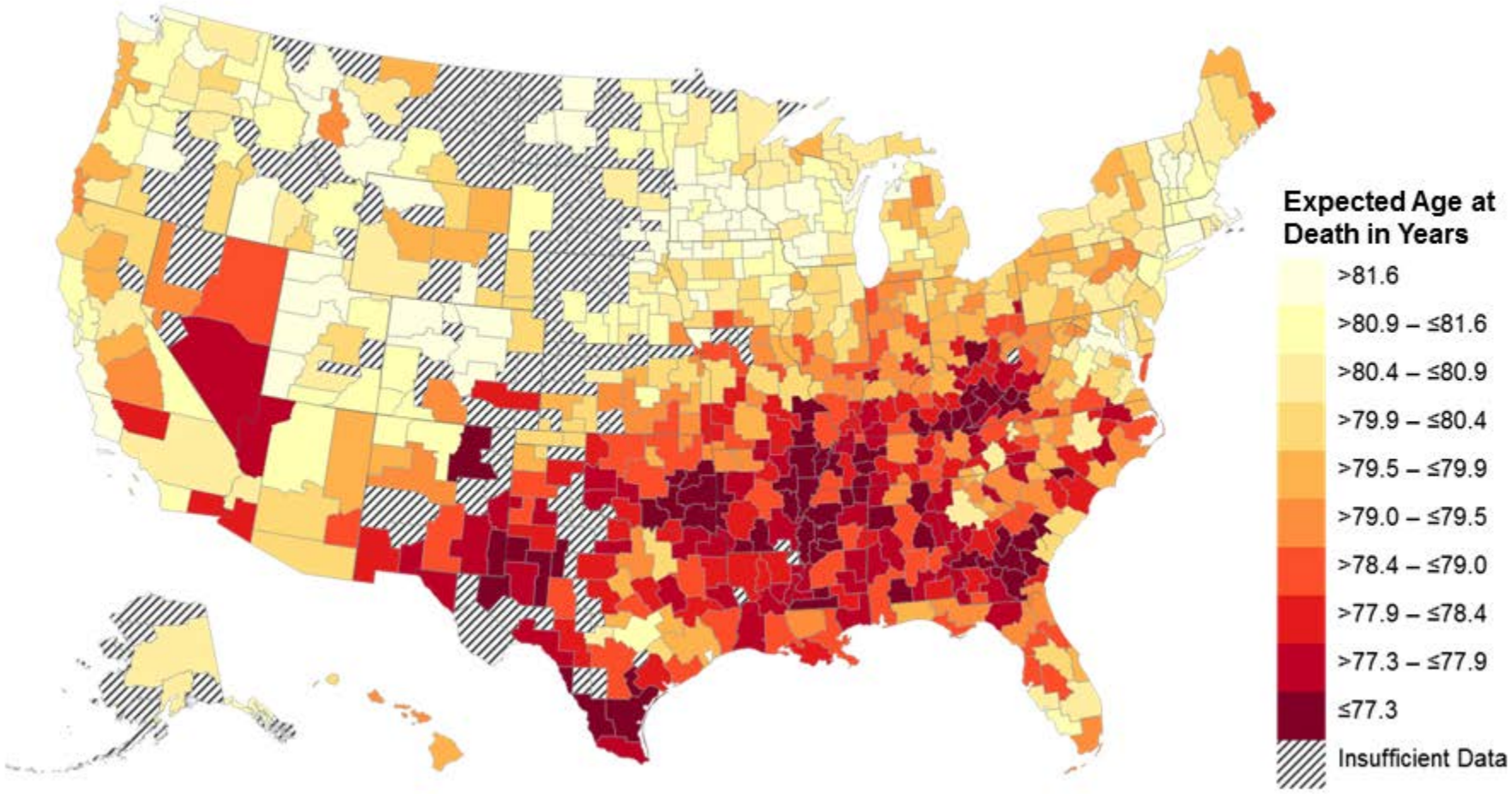


Expected Age at Death for 40 Year Old Men
Bottom Quartile of U.S. Income Distribution



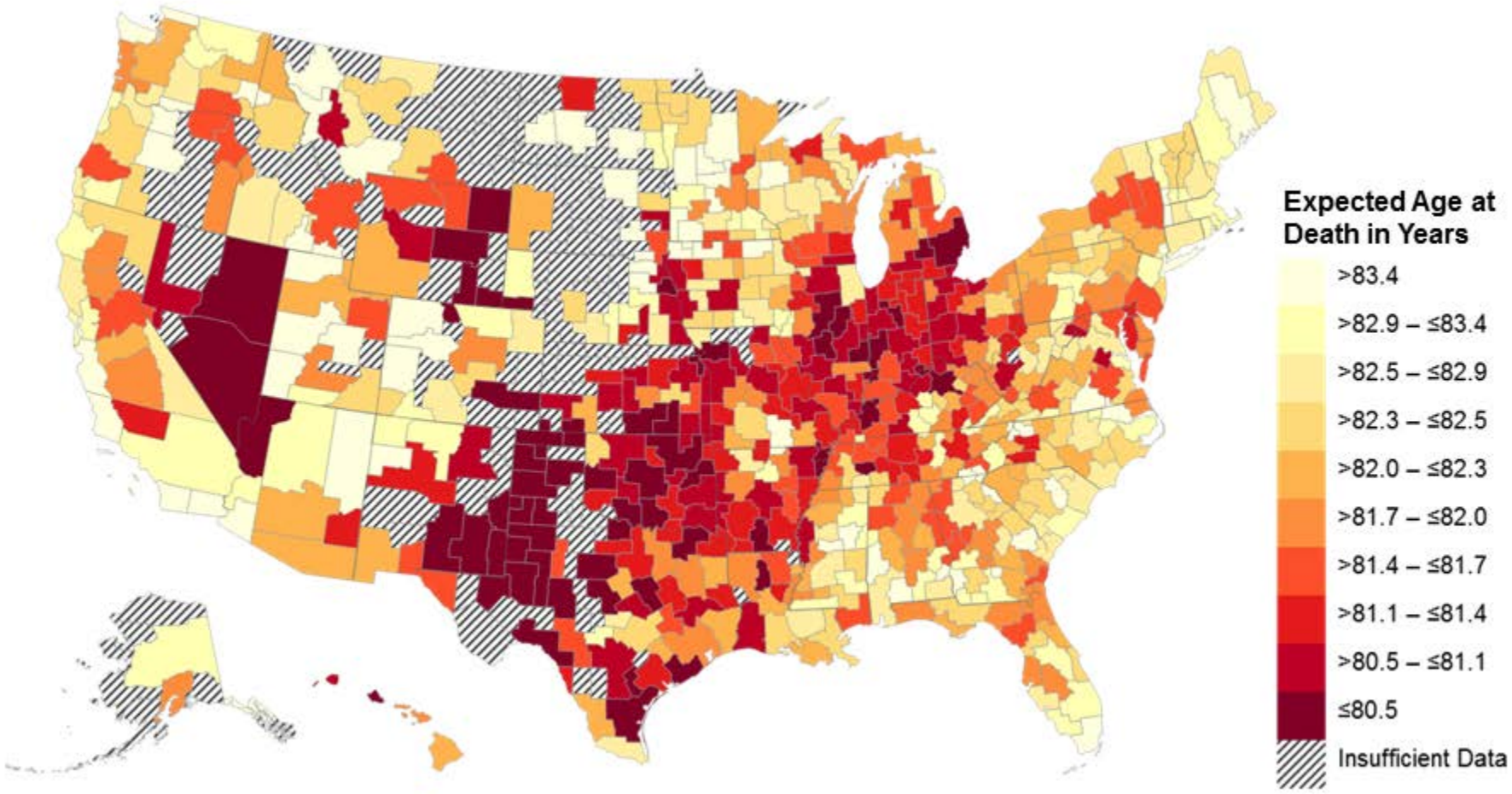
Note: Lighter Colors Represent Areas with Higher Life Expectancy

Expected Age at Death for 40 Year Old Men
Pooling All Income Groups



Note: Lighter Colors Represent Areas with Higher Life Expectancy

Expected Age at Death for 40 Year Old Women
Bottom Quartile of U.S. Income Distribution



Note: Lighter Colors Represent Areas with Higher Life Expectancy

Expected Age at Death for 40 Year Olds in Bottom Quartile
Top 10 and Bottom 10 CZs Among 100 Largest CZs

Top 10 CZs			Bottom 10 CZs		
Rank	CZ	Expected Age at Death	Rank	CZ	Expected Age at Death
1	New York, NY	81.8 (81.6, 82.0)	91	San Antonio, TX	78.0 (77.6, 78.4)
2	Santa Barbara, CA	81.7 (81.3, 82.1)	92	Louisville, KY	77.9 (77.7, 78.2)
3	San Jose, CA	81.6 (81.2, 82.0)	93	Toledo, OH	77.9 (77.6, 78.2)
4	Miami, FL	81.2 (80.9, 81.6)	94	Cincinnati, OH	77.9 (77.7, 78.1)
5	Los Angeles, CA	81.1 (80.9, 81.4)	95	Detroit, MI	77.7 (77.5, 77.8)
6	San Diego, CA	81.1 (80.8, 81.4)	96	Tulsa, OK	77.6 (77.4, 77.9)
7	San Francisco, CA	80.9 (80.6, 81.3)	97	Indianapolis, IN	77.6 (77.4, 77.8)
8	Santa Rosa, CA	80.8 (80.5, 81.2)	98	Oklahoma City, OK	77.6 (77.3, 77.8)
9	Newark, NJ	80.7 (80.5, 80.9)	99	Las Vegas, NV	77.6 (77.4, 77.8)
10	Port St. Lucie, FL	80.7 (80.5, 80.9)	100	Gary, IN	77.4 (77.1, 77.8)

Note: 95% confidence intervals shown in parentheses

Why Does Life Expectancy for Low-Income Individuals Vary Across Areas?

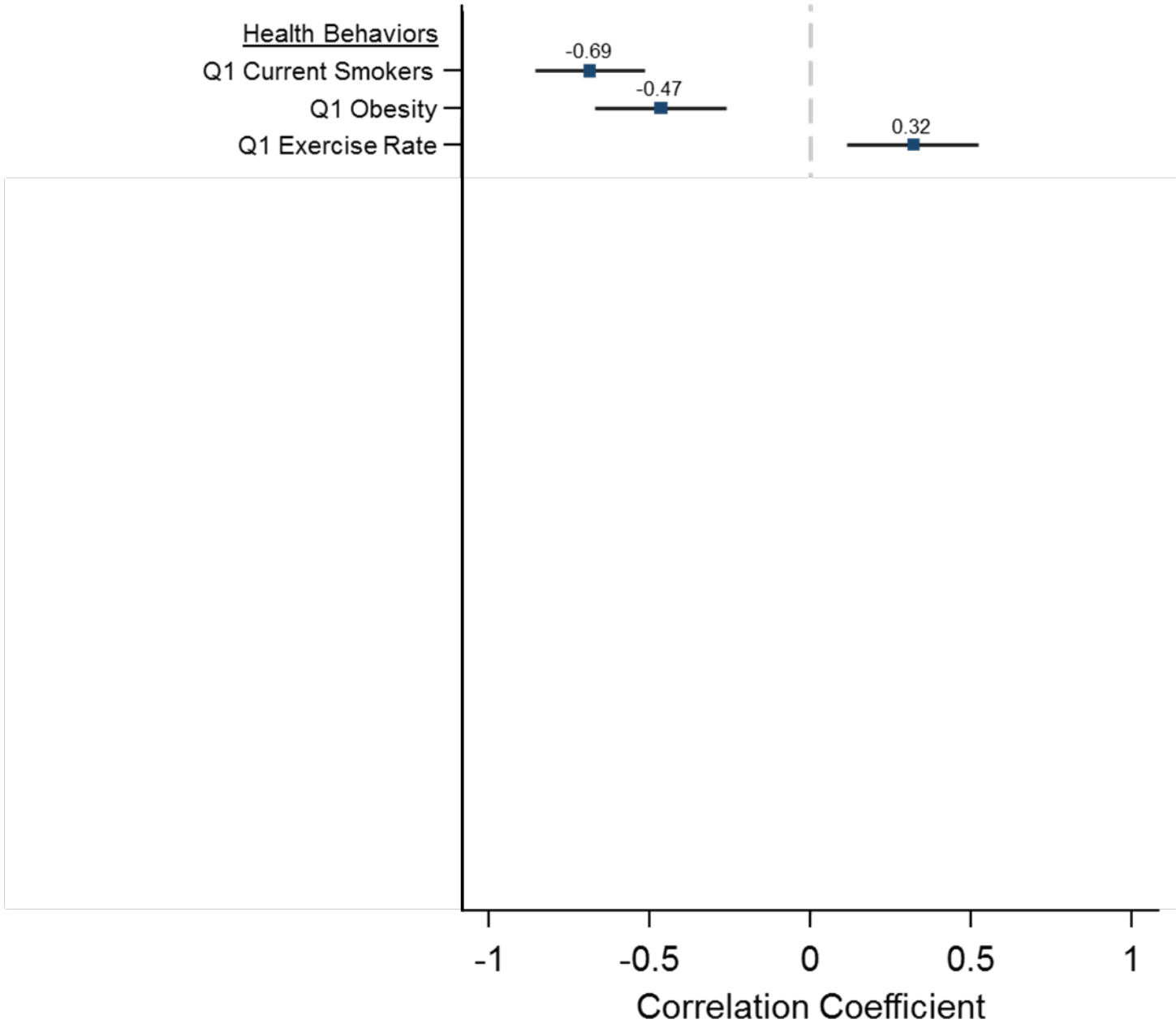
Why Does Life Expectancy for Low-Income Individuals Vary Across Areas?

- Now use local area variation to explore determinants of life expectancy

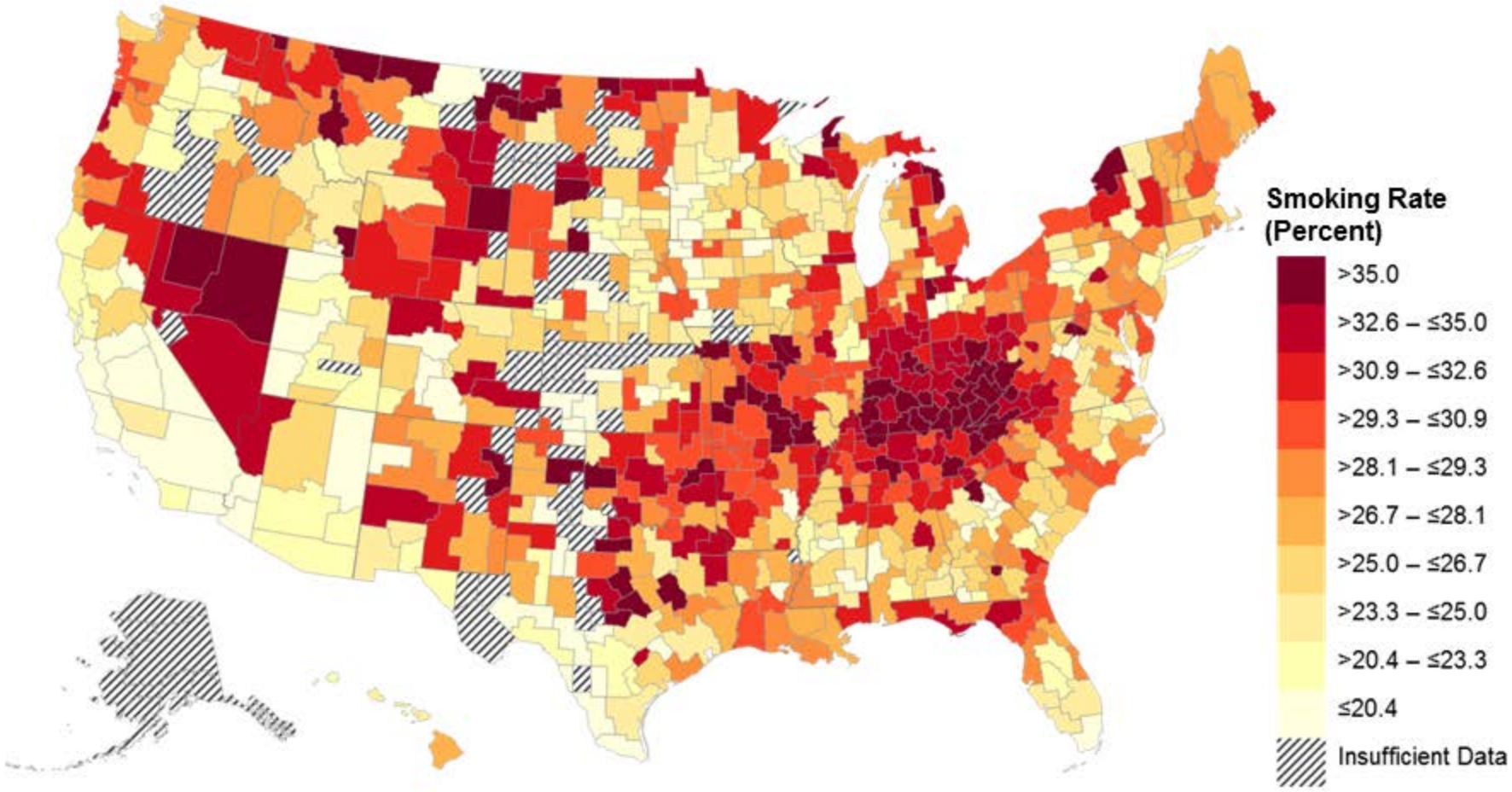
- Key question: is lower life expectancy in some areas driven by lack of access to health care or differences in health behavior?

- Correlate life expectancy estimates with measure of health care access and health behaviors to answer this question

Correlations of Expected Age at Death with Health and Social Factors For Individuals in Bottom Quartile of Income Distribution

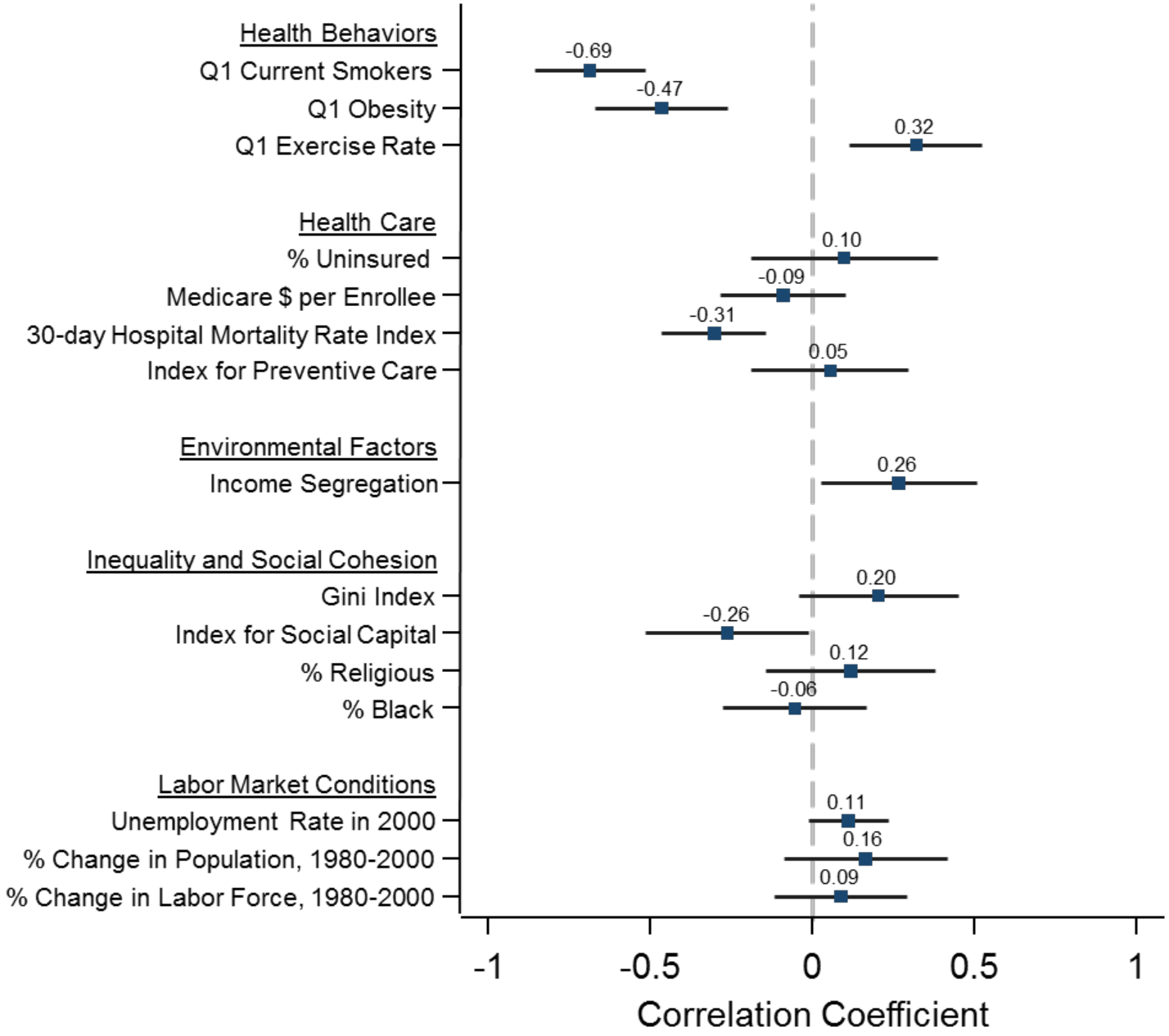


Smoking Rates for Individuals in Bottom Income Quartile

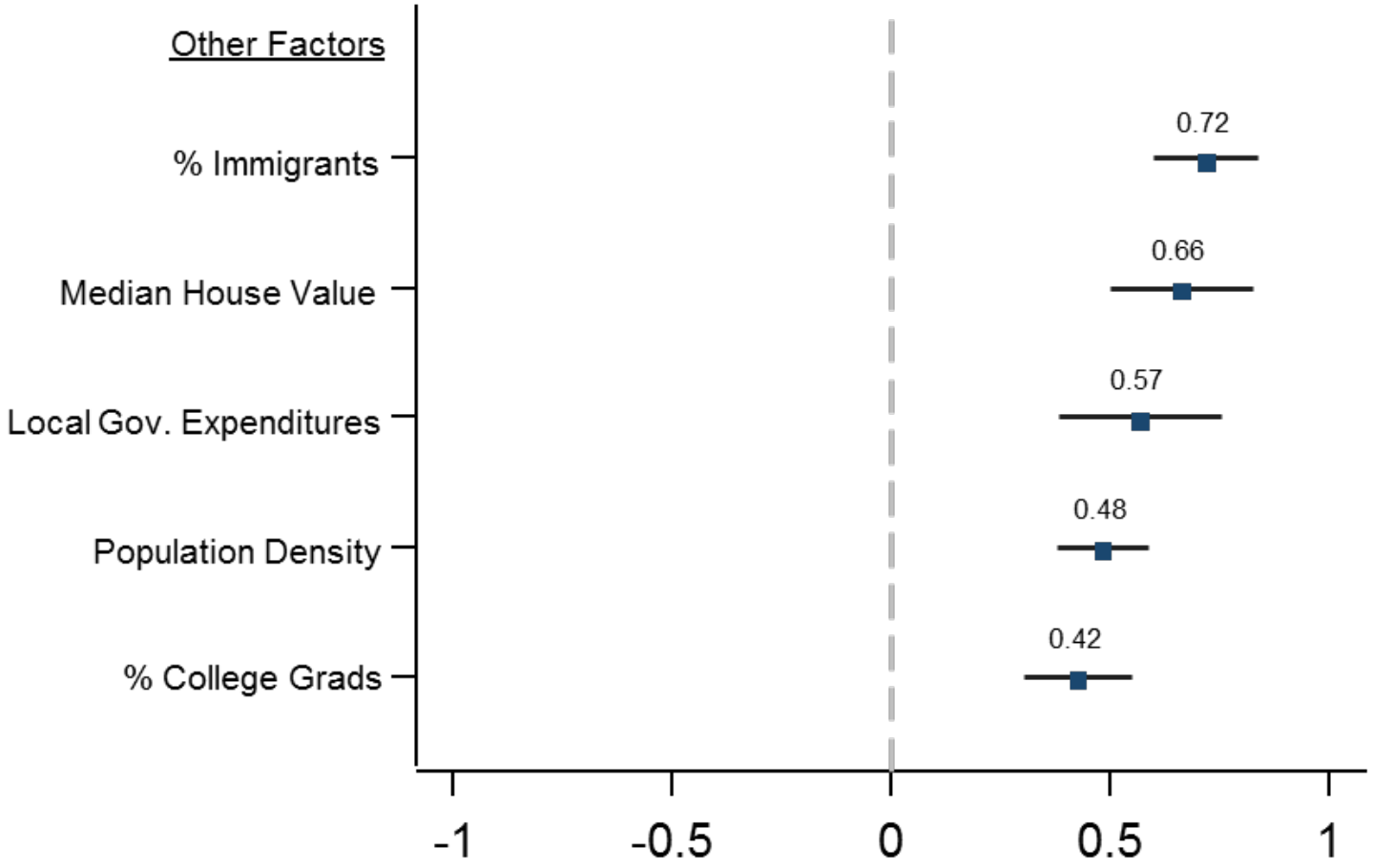


Note: Lighter Colors Represent Areas Lower Smoking Rates

Correlations of Expected Age at Death with Health and Social Factors For Individuals in Bottom Quartile of Income Distribution



Correlations of Expected Age at Death with Other Factors For Individuals in Bottom Quartile of Income Distribution



Why Does Life Expectancy for Low-Income Individuals Vary Across Areas?

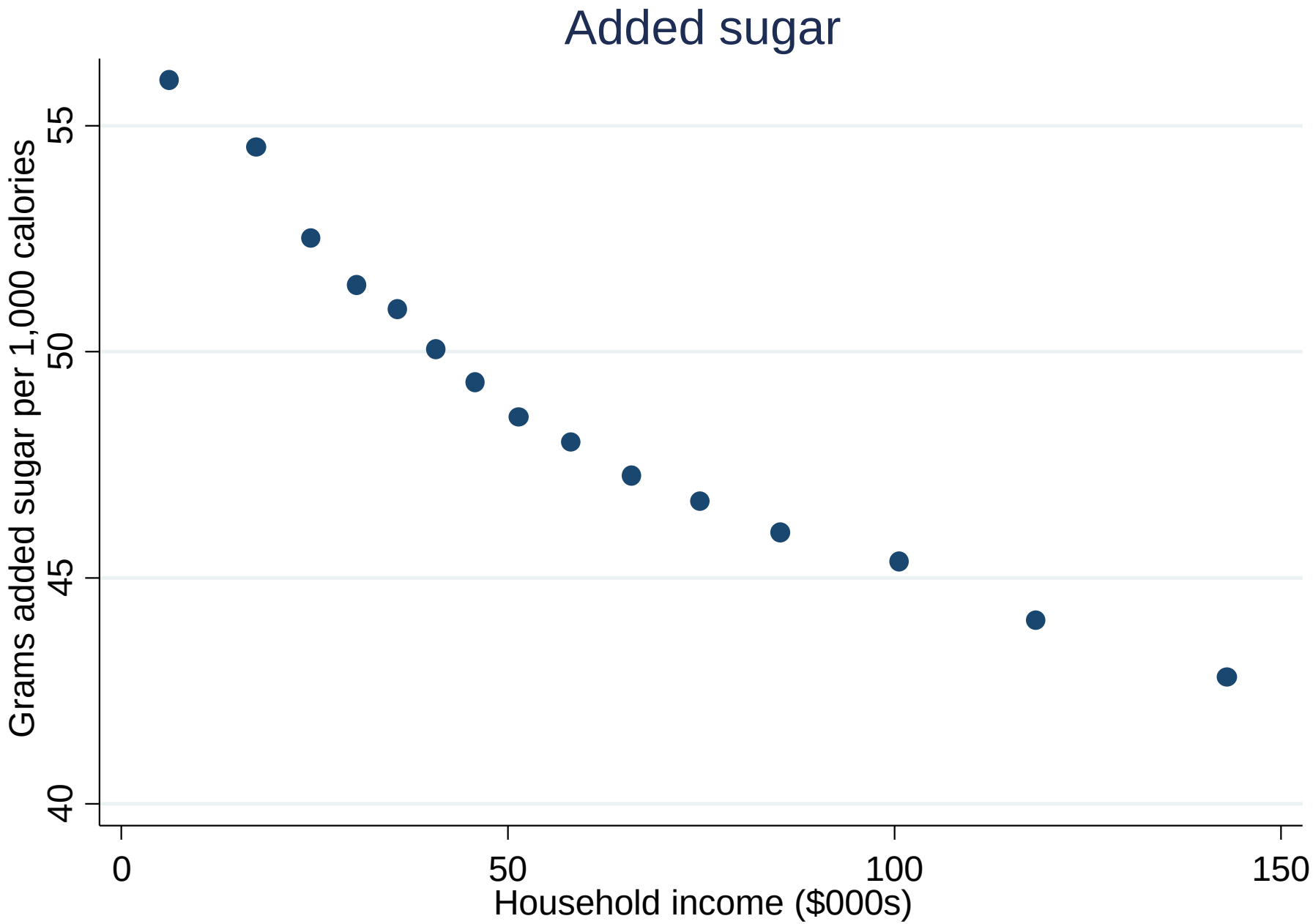
- Local area variation suggests that differences in health behaviors are more predictive of life expectancy than differences in health care access

- Further evidence for this view comes from directly examining nutritional patterns

Differences in Nutrition by Income

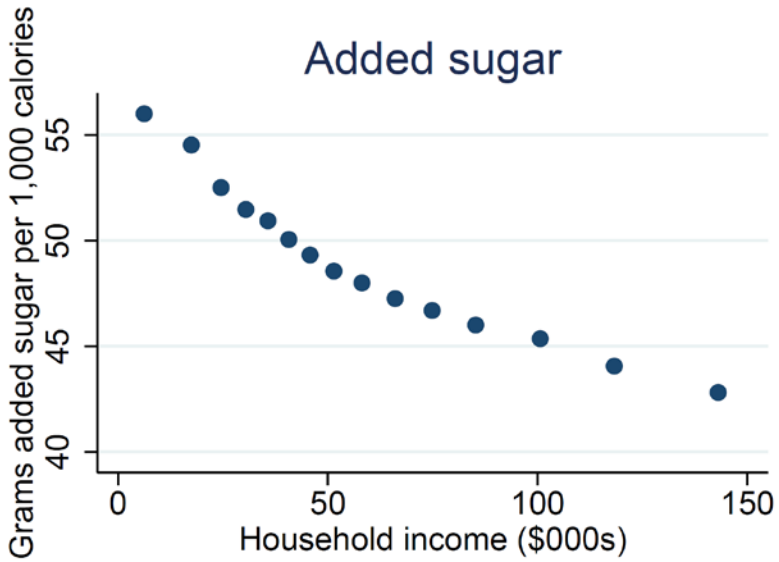
- Alcott et al. (2018) use Nielsen homescan data on grocery store purchases to examine how nutrition varies with income
  - About 170,000 households who scan all of their purchases and record UPCs, which are then matched to nutritional information from the USDA

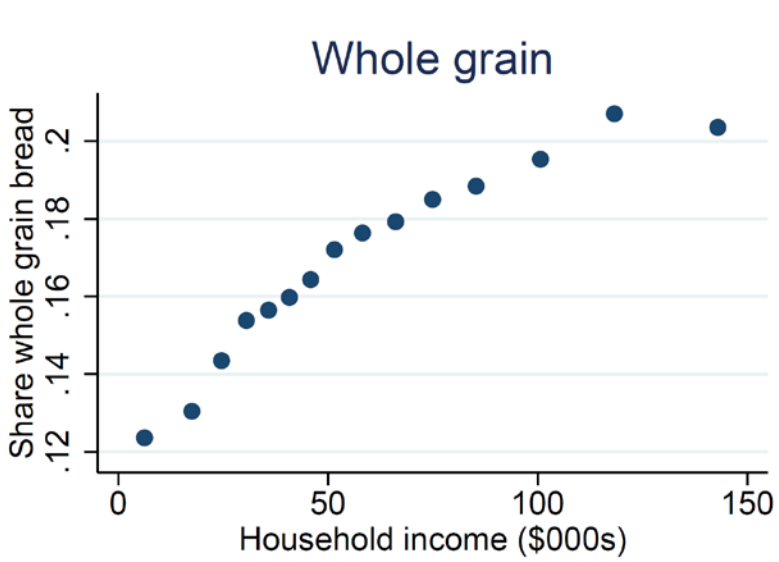
Healthfulness of Grocery Purchases by Household Income

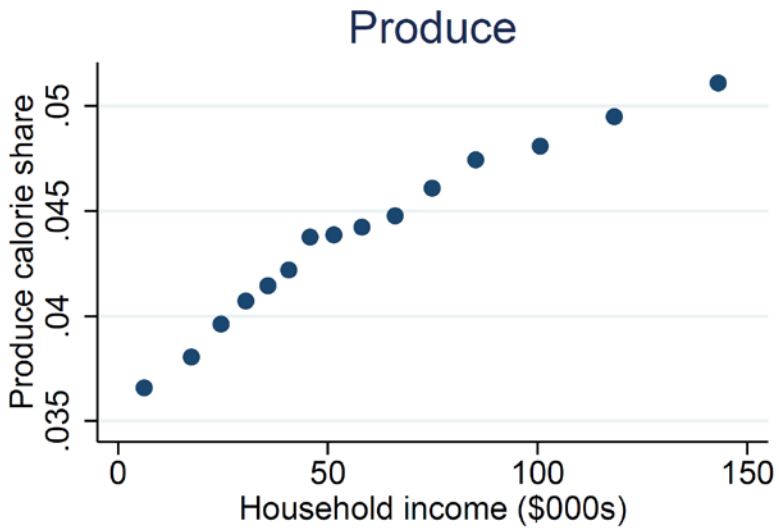


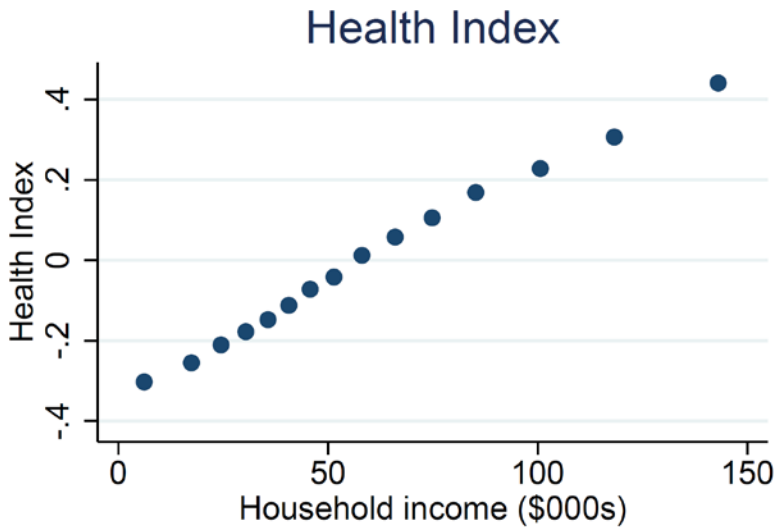
Source: Allcott, Diamond, Dube, Handbury, Rahkovsky, and Schnell 2018

Healthfulness of Grocery Purchases by Household Income







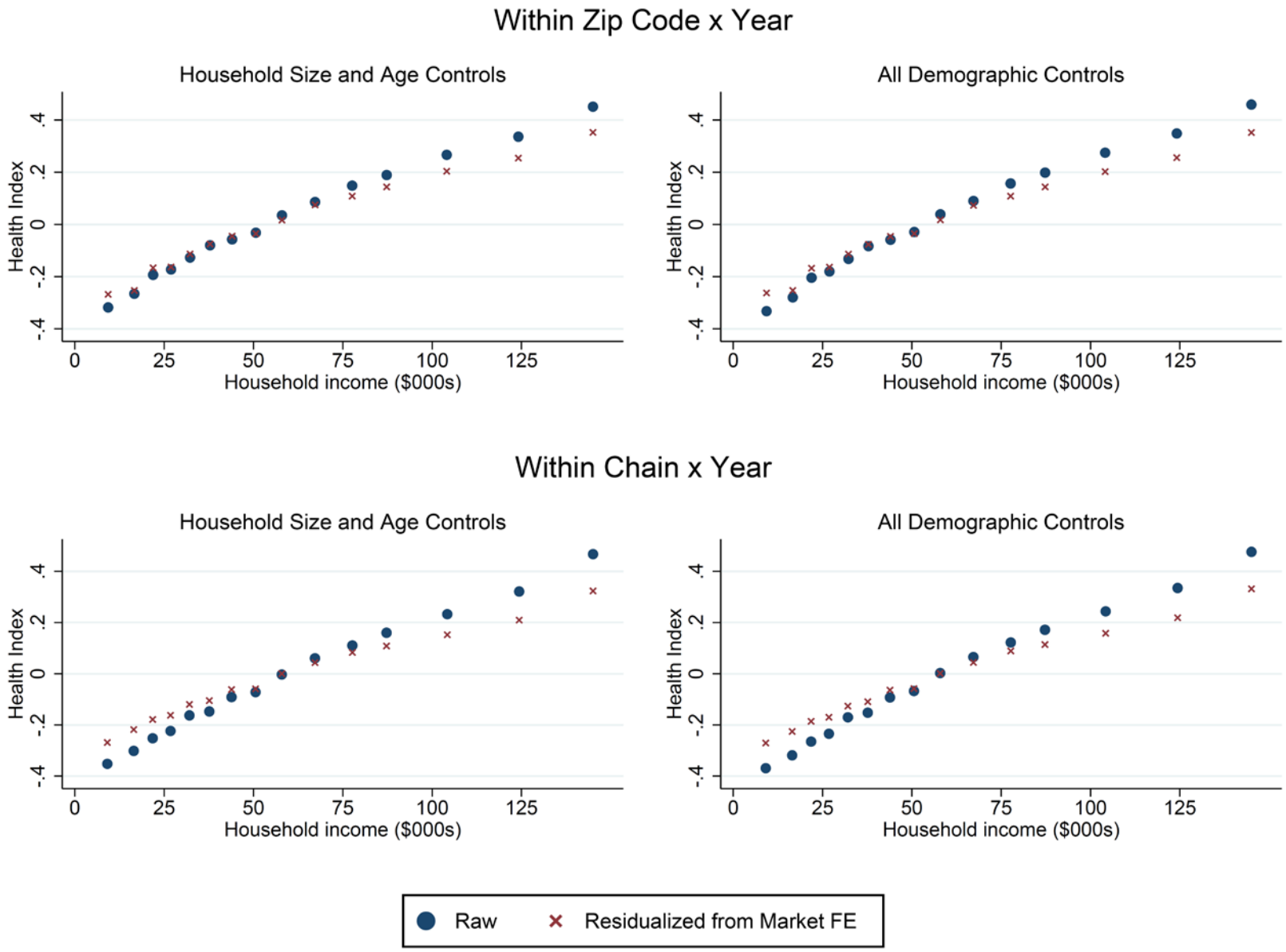


Source: Allcott, Diamond, Dube, Handbury, Rahkovsky, and Schnell 2018

Differences in Nutrition by Income

- These differences in nutrition are *not* driven by a lack of access to health food (“food deserts”)

Healthfulness of Grocery Purchases by Household Income that Shop in the Same Market



Source: Allcott, Diamond, Dube, Handbury, Rahkovsky, and Schnell 2018

Differences in Health Behaviors by Income

- These differences in nutrition are *not* driven by a lack of access to health food (“food deserts”)

- Again suggests that differences in health outcomes are not caused by a direct lack of access to resources

  - Instead, appear to be due to different *choices* made by lower-income households

Differences in Health Behaviors by Income

- Why do low income households tend to have less healthy behaviors?

- One hypothesis: effects of environment and resources at early ages on preferences
  - Ex: Atkin (2016) studies migrants in India and shows that nutritional habits formed at young ages persist for many years after people move

- Alternative hypothesis: lack of income constrains choice (unhealthy foods may be less expensive per calorie)
  - Discuss this economic explanation in next lecture with Jesse Shapiro